The Fourier-Galerkin Method for Band Structure Computations of 2D and 3D Photonic Crystals

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

M. Sc. Branimir Anić
aus Balingen

Tag der mündlichen Prüfung: 11. Dezember 2013

Referent:	Prof. Dr. Willy Dörfler
Koreferentin:	Prof. Dr. Marlis Hochbruck

Acknowledgements

This dissertation emerged during my work as a research and teaching assistant at the Institute for Applied Mathematics and Numerical Analysis of the Karlsruhe Institute of Technology (KIT). The existing research program of the DFG Research Training Group 1294 "Analysis, Simulation and Design of Nanotechnological Processes" at the Department of Mathematics gave me the possibility to choose a fascinating and challenging topic for my research, which was embedded into the Research Training Group 1294.

I would like to take the opportunity to thank several people for their support during my research. First, I would like to thank my advisors, Prof. Dr. Willy Dörfler and Prof. Dr. Marlis Hochbruck. I appreciate their support and their willingness to discuss my problems whenever I needed to. I am grateful to both that they always made me feel welcome for discussions concerning my research.

I would also like to thank my colleagues at the Institute for Applied Mathematics and Numerical Analysis. Foremost, I would like to mention Dr. Markus Richter and Markus Maier. I am very thankful to both for the discussions and the proofreading. Moreover, I would like to thank Dr. Markus Richter for his permission to use the Figures 6.1, 6.2, 6.3, 8.1 and 8.2.

The existence of the Research Training Group 1294 Analysis, Simulation and Design of Nanotechnological Processes at the KIT gave me the opportunity to work on an interesting and challenging topic. Moreover, the financial support by the Research Training Group gave me the possibility to attend interesting conferences and workshops.

I am very grateful for the financial support of the Karlsruhe House of Young Scientists under the KHYS Networking Grant.

Finally, I want to thank my family for their persistent encouragement. I am especially indebted to my wonderful wife Marijana who was a tremendous support for me during the last years.

Branimir Anić
Karlsruhe, November 2013

Contents

Chapter 1

Introduction

1.1 Motivation

Many phenomena in physics and the natural sciences can be described mathematically by partial differential equations (PDEs). There exist many excellent textbooks on the topic of PDEs, for example [25, 54, 72], where for different types of PDEs the question about existence and uniqueness of solutions is being discussed. In general it is not possible to find an explicit representation of a solution to a PDE. Therefore, various methods for the numerical approximation to the unknown solutions of PDEs have been developed in the past decades, e.g. the Finite Differences Method [67], the Finite Element Method [3, 8, 66], the Boundary Element Method [59] and Fourier methods [11, 29, 52].

In this thesis we consider the propagation of light waves in *photonic crystals*, which are an important class of physical structures investigated in *nanotechnology*. Photonic crystals are materials, which are composed of at least two different dielectrics or metals, and which exhibit a spatially periodic structure, typically at the length scale of a few hundred nanometers [24]. Depending on whether the periodicity extends into one, two or three space dimensions, a photonic crystal is called one-, two- or three-dimensional [55]. The goal is to develop new data transmission and processing device concepts like optical computers with the help of three-dimensional photonic crystals. However, the manufacturing of photonic crystals is still expensive, and so far no large scale production methods could be devised. This is why numerical simulation is a very important tool in this area.

An important property of photonic crystals is that their periodic nanostructure affects the propagation of light waves at certain optical frequencies. An incident light wave is subject to periodic, multiple diffraction, resulting in coherent wave interference inside the crystal [55]. Depending on the frequency of the light wave, this interference can be of destructive nature. In this case the propagation of the

light wave inside the crystal is not permitted. Depending on the spatial distribution of the material, for a certain range of optical wave frequencies the propagation of light is allowed or not. In the latter case, where propagation is not permitted, it is common to name this frequency range a *photonic band gap*. An introduction to this topic can be found in [24, 34].

1.2 Goals

The partial differential equations which are used for the mathematical modeling of photonic crystals are the *Maxwell equations* with spatially periodic permittivities. The aim of this thesis is the mathematical analysis and numerical implementation of *photonic band structure* (this will be explained in Section 3.5) computations for 2D and 3D photonic crystals with the *Fourier-Galerkin method*. There are many other discretization methods that are widely accepted and analyzed in the mathematics community for these kinds of problems, especially the *Finite Element method*. Discretization with Fourier methods for these types of problems is widespread among physicists and engineers. However, there is nearly no theoretical convergence analysis for Fourier methods used in photonics. The first work in which the convergence of the Fourier-Galerkin method for the one-dimensional case was studied is [45]. The only known work related to photonics, which deals with convergence analysis for this kind of problem is [52]. In that work Norton and Scheichl consider the computation and analysis of the spectrum of a 2D Schrödinger operator with a periodic potential. In this thesis we want to study 2D and 3D benchmark problems which have also been considered in [10, 16, 17, 22, 24, 39, 55, 64]. Our goal is to analyze the implementation and the convergence of 2D and 3D photonic band structure computations with the Fourier-Galerkin method. We will see that for the *Helmholtz equation*, that arises in the 2D situation, this method has desirable properties. Moreover, we will discuss the discretization and implemtenation for the 3D problem. We will see that this problem can be solved without any preconditioning with the *Harmonic Restarted Arnoldi* algorithm.

1.3 Outline

This thesis is organized as follows. In Chapter 2 we introduce the notations and conventions that will be frequently used in this work. Moreover, we will introduce the function spaces that will be needed in the following chapters. In Chapter 3 we present the mathematical model for wave propagation in photonic crystals. After having introduced the Maxwell equations, we discuss periodic structures and the resulting eigenvalue problems for band structure computations. In Chapter 4 we

discuss how elliptic boundary value problems can be discretized with the Fourier-Galerkin method. Moreover, we will explain how for the arising matrix structures matrix-vector products can be realized efficiently via FFT, and thus can be used for iterative methods. Li and Haggans discussed in [42] convergence problems that can arise when for discontinuous structures Fourier methods are used for the numerical approximation of eigensolutions to the Maxwell equations. In [40] it was shown that a reformulation of the problem yields better convergence rates. In [41] Li stated theorems about which Fourier factorization in which setting has to be used, and thus explained why the reformulation in [40] works. Five years later, the proof of Li's Fourier factorization theorems was presented in the appendix of Chapter 4 in [7]. In Chapter 5 we will discuss Li's theorems and their proofs in much more detail. This chapter will be self-contained, therefore we will reiterate all the definitions from [7] and [41]. Moreover, we will treat the proofs in great detail and present all tools that are needed for the proofs. We will also consider several examples that illustrate the theorems. In Chapter 6 we will discuss the discretization of a scalar Helmholtz and a divergence-type problem for a 2D photonic crystal. We will see that this problem can be solved efficiently on a usual desktop PC. After that we will analyze the convergence for the Helmholtz problem, similarly as it was done in [52]. The discretization of the vector valued 3D Maxwell problem will be treated in Chapter 7. The underlying numerical linear algebra techniques for the solution of the discretized problem will be discussed in Chapter 8. Conclusions and an outlook on future work are given in Chapter 9.

Chapter 2

Preliminaries

In this chapter we introduce notations, conventions and the function spaces permanently used in the following chapters.

2.1 Notations and conventions

We start with conventions and notations that will be used throughout this work. First we consider notations for scalars and vectors. When a boldface symbol occurs this will indicate that we are dealing with a vectorial quantity, in contrast to a scalar one. For example, a vector $\boldsymbol{x} \in \mathbb{R}^d$ is represented by a boldface letter in contrast to a scalar $a \in \mathbb{R}$, which is represented by a non-boldface letter. Matrices will always be denoted by capital letters, e.g. $A \in \mathbb{C}^{m\times n}$. If we write A^{H}, then we mean the conjugate transpose of a matrix $A \in \mathbb{C}^{m\times n}$, i.e. $A^{\mathrm{H}} \in \mathbb{C}^{n\times m}$ with

$$A^{\mathrm{H}} = \overline{A}^{\top}.$$

When we consider sums we will often use the *index range* $\mathbb{I}_N$, which is defined in the following way:

$$\mathbb{I}_N := \left\{\boldsymbol{n} \in \mathbb{Z}^d \text{ with } \|\boldsymbol{n}\|_\infty \leq N\right\} \tag{2.1}$$

for $N \in \mathbb{N}$. For those index ranges depending on $N \in \mathbb{N}$ we always choose *lexicographic ordering* when they are used in a summation, i.e. in the 2D case this means that we run through the set $\mathbb{I}_N = [-N, ..., N]^2$ by means of

$$\begin{pmatrix}-N\\-N\end{pmatrix}, \dots, \begin{pmatrix}N\\-N\end{pmatrix}, \begin{pmatrix}-N\\-N+1\end{pmatrix}, \dots, \begin{pmatrix}N\\-N+1\end{pmatrix}, \dots, \begin{pmatrix}-N\\N\end{pmatrix}, \dots, \begin{pmatrix}N\\N\end{pmatrix},$$

and in the 3D case

$$\begin{pmatrix}-N\\-N\\n\end{pmatrix}, \dots, \begin{pmatrix}N\\-N\\n\end{pmatrix}, \begin{pmatrix}-N\\-N+1\\n\end{pmatrix}, \dots, \begin{pmatrix}N\\-N+1\\n\end{pmatrix}, \dots, \begin{pmatrix}-N\\N\\n\end{pmatrix}, \dots, \begin{pmatrix}N\\N\\n\end{pmatrix},$$

for n from $-N$ to N. So if we are in the 2D case, then for example the ordering for $\mathbb{I}_1$ is

$$\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}0\\-1\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}-1\\0\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}.$$

In the context of matrices, the symbol I_N will always denote the $N \times N$ *identity matrix*. If we are working with vectorial quantities, then by $\boldsymbol{e}_j$ we denote the j-th *unit vector*. For any two matrices $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$ their *Kronecker product* $A \otimes B$ is defined as the $mp \times nq$ matrix

$$A \otimes B := \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}.$$

Next we want to define the **vec** *operator*. If we have a matrix $A \in \mathbb{C}^{m\times n}$ we can represent it as an object with n column vectors of length m, namely $A = [\boldsymbol{a}_1 \cdots \boldsymbol{a}_n]$. Then the **vec** operator creates a column vector by stacking the columns of A below one another:

$$\mathbf{vec}(A) := \begin{pmatrix} \boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}.$$

For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3$ the *cross product* $\boldsymbol{a} \times \boldsymbol{b}$ is defined as

$$\boldsymbol{a} \times \boldsymbol{b} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} := \begin{pmatrix} a_2b_3 - a_3b_2 \\ a_3b_1 - a_1b_3 \\ a_1b_2 - a_2b_1 \end{pmatrix}.$$

The cross product $\boldsymbol{a} \times \boldsymbol{b}$ of two vectors can be represented as a matrix-vector product $A\boldsymbol{b}$. The matrix A that acts on the vector $\boldsymbol{b} \in \mathbb{R}^3$ is given by

$$A := \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix}. \tag{2.2}$$

Following the definition of the cross product we can define the *curl operator* for three-dimensional vector fields $\boldsymbol{f}$. We define

$$\nabla\times \boldsymbol{f} := \begin{pmatrix} \frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z} \\ \frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x} \\ \frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \end{pmatrix}.$$

Next, we define the so-called *Hadamard product* of two matrices $A, B \in \mathbb{C}^{m\times n}$. The Hadamard product $A \odot B$ is again an $m \times n$ matrix and is defined as

$$(A \odot B)_{ij} := a_{ij} b_{ij}.$$

We will also call this operation the *pointwise product*. Later this will be useful, because the multiplication of a diagonal matrix with a vector can be represented as a pointwise product of two vectors.

In general, computing a matrix-vector product $A\boldsymbol{x}$ for $A \in \mathbb{R}^{N\times N}$ and $\boldsymbol{x} \in \mathbb{R}^N$ has the compuational cost of $\mathcal{O}(N^2)$ operations. However, if a matrix has a special structure (as we will have later in our discretizations), then this cost can be reduced to $\mathcal{O}(N \log(N))$ operations. We will explain in section 4.2 how this can be achieved. Now we want to introduce this special class of matrices. A *Toeplitz matrix* $T = (t_{ij})$ is a matrix whose entries t_{ij} only depend on $i - j$. This means that a Toeplitz matrix $T \in \mathbb{R}^{N\times N}$ has the same entries along all diagonals parallel to the principal diagonal, i.e. T is of the form

$$T = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \dots & \dots & t_{1-N} \\ t_1 & t_0 & t_{-1} & \ddots & & t_{2-N} \\ t_2 & t_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & t_{-1} & t_{-2} \\ \vdots & & \ddots & t_1 & t_0 & t_{-1} \\ t_{N-1} & \dots & \dots & t_2 & t_1 & t_0 \end{pmatrix}. \tag{2.3}$$

In the following chapters we will use the expression $[\![f]\!]$ which denotes a Toeplitz matrix generated by the Fourier coefficients $\widehat{f}_k$, $k \in \mathbb{Z}$, of some $\mathbb{Z}$-periodic function $f : \mathbb{R} \to \mathbb{R}$. Such a matrix is of the form

$$[\![f]\!] := \begin{pmatrix} \widehat{f}_0 & \widehat{f}_{-1} & \widehat{f}_{-2} & \dots & \dots & \widehat{f}_{-2N} \\ \widehat{f}_1 & \widehat{f}_0 & \widehat{f}_{-1} & \ddots & & \widehat{f}_{1-2N} \\ \widehat{f}_2 & \widehat{f}_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \widehat{f}_{-1} & \widehat{f}_{-2} \\ \vdots & & \ddots & \widehat{f}_1 & \widehat{f}_0 & \widehat{f}_{-1} \\ \widehat{f}_{2N} & \dots & \dots & \widehat{f}_2 & \widehat{f}_1 & \widehat{f}_0 \end{pmatrix}. \tag{2.4}$$

For the definition of Fourier coefficients see Section 2.2.3. If a matrix has the form as in (2.3), and the entries t_j, $j = 1-N, ..., N-1$, themselves are Toeplitz matrices, then we call this a *block Toeplitz matrix with Toeplitz blocks* (BTTB). Another important class of matrices are the so-called *circulant matrices*. A circulant matrix

is completely determined by its first column and is of the form

$$C = \begin{pmatrix} c_0 & c_{N-1} & \dots & c_2 & c_1 \\ c_1 & c_0 & \ddots & c_3 & c_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{N-2} & c_{N-3} & \ddots & c_0 & c_{N-1} \\ c_{N-1} & c_{N-2} & \dots & c_1 & c_0 \end{pmatrix}. \tag{2.5}$$

This means that for a circulant matrix the entries of each column are the same as in the previous one. A column is built by shifting down the entries of the previous column, and taking the last element as the first. Circulant matrices are a special class of Toeplitz matrices, i.e. every circulant matrix is a Toeplitz matrix but not vice versa. If a matrix has the form as in (2.5), and the entries c_j, $j = 0, ..., N-1$, itself are circulant matrices, then we call the resulting matrix a *block ciculant matrix with ciculant blocks* (BCCB). For a general treatment of circulant matrices see [19].

2.2 Function spaces

In this section we want to introduce the function spaces that will be needed in the following chapters. We will only collect the results and tools that we need. For a more complete discussion we refer for Lebesgue spaces to [57], for Sobolev spaces to [1, 2, 8, 9, 25, 54] and for trigonometric function spaces as well as Fourier series to [5, 70]. All the tools that we present here can be found in much greater detail in those textbooks.

2.2.1 Lebesgue spaces

In this section we introduce the so-called *Lebesgue spaces* L^p. Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ and $u : \Omega \to \mathbb{R}$ a measurable function. For $p \in [1, \infty)$ we define the norm

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u(\boldsymbol{x})|^p d\boldsymbol{x} \right)^{\frac{1}{p}},$$

and for $p = \infty$ we define

$$\|u\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{\boldsymbol{x} \in \Omega} |u(\boldsymbol{x})|.$$

The Lebesgue spaces $L^p(\Omega)$, for $p \in [1, \infty]$ are defined as

$$L^p(\Omega) := \{u : \Omega \to \mathbb{R} \,:\, u \text{ measurable and } \|u\|_{L^p(\Omega)} < \infty\}.$$

The space $L^2(\Omega)$ will be of special interest in the following chapters. It is the space of square integrable functions. With the *L^2-inner product*

$$(f,g) := \int_\Omega f(\boldsymbol{x})\overline{g(\boldsymbol{x})}d\boldsymbol{x}$$

we obtain

$$\|u\|^2_{L^2(\Omega)} = (u,u) = \int_\Omega |u(\boldsymbol{x})|^2 d\boldsymbol{x}.$$

Whenever we use the inner product $(\cdot,\cdot)$ in this work, the L^2-inner product is meant. For the vectorial setting we define $\boldsymbol{L}^2(\Omega) := L^2(\Omega)^3$. Actually, the spaces $L^p(\Omega)$ are equivalence classes of functions where functions are identified which only differ on sets of measure zero. In the next theorem we present the so-called *Hölder inequality*.

Theorem 2.2.1 ([57], Theorem 3.8)**.** *Let $1 \le p,q \le \infty$ sucht that $\frac{1}{p}+\frac{1}{q}=1$. If $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$, then $uv \in L^1(\Omega)$ and*

$$\|uv\|_{L^1(\Omega)} \le \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)}.$$

2.2.2 Sobolev spaces

In this section we introduce the so-called *Sobolev spaces*, which have emerged as the right spaces for analyzing PDEs. First we want to introduce a generalized concept of differentiation. We define the set of *locally integrable* functions on Ω by

$$L^1_{\text{loc}}(\Omega) := \left\{u : \Omega \to \mathbb{R} \,:\, u \in L^1(K) \text{ for all } K \subset\subset \Omega\right\}$$

and the space of *test functions* by

$$C_0^\infty(\Omega) := \{\phi \in C^\infty(\Omega) \,:\, \operatorname{supp}(\phi) \subset \Omega\}$$

We call a vector $\alpha \in \mathbb{N}_0^d$ a d-dimensional *multi-index* with the corresponding *order*

$$|\alpha| = \alpha_1 + ... + \alpha_d.$$

Then the α-th partial derivative $D^\alpha u$ is defined by

$$D^\alpha u := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} u.$$

Now suppose that $u, v \in L^1_{\text{loc}}(\Omega)$ and α is a multi-index. We say that u is *weakly differentiable*, with the α-th *weak partial derivative* v, if

$$\int_\Omega u D^\alpha \phi \, d\boldsymbol{x} = (-1)^\alpha \int_\Omega v\phi \, d\boldsymbol{x}$$

holds for all test functions $\phi \in C_0^\infty(\Omega)$. If u is weakly differentiable, then its *weak derivative* v is uniquely determined. In this case we simply write $D^\alpha u = v$. Now we are ready to define the Sobolev spaces $W^{m,p}(\Omega)$. For a fixed $p \in [1, \infty]$ and $m \in \mathbb{N}$ we define

$$W^{m,p}(\Omega) := \{u \in L^p(\Omega) \,:\, D^\alpha u \in L^p(\Omega) \text{ for all } |\alpha| \leq m\}.$$

For the special case $p = 2$ it is common to write $H^2(\Omega) = W^{m,2}(\Omega)$. The letter H is used in honor of David Hilbert, in order to symbolize that it is a Hilbert space. Finally we want to introduce *fractional Sobolev spaces*, also called *Sobolev-Slobodeckij spaces*. Let $s \in \mathbb{R}_{\geq 0}$ and $\lfloor s \rfloor := \max\{m \in \mathbb{N} : m \leq s\}$. Then we define

$$W^{s,p}(\Omega) := \left\{u : \Omega \to \mathbb{R} \,:\, \|u\|_{W^{s,p}(\Omega)} < \infty\right\},$$

where

$$\|u\|^p_{W^{s,p}(\Omega)} := \|u\|^p_{W^{m,p}(\Omega)} + \sum_{|\alpha|=m} \int_\Omega \int_\Omega \frac{|D^\alpha u(\boldsymbol{x}) - D^\alpha u(\boldsymbol{y})|^p}{|\boldsymbol{x} - \boldsymbol{y}|^{(s-m+d/p)p}} d\boldsymbol{x} d\boldsymbol{y}$$

with $m := \lfloor s \rfloor$. Later, in Section 6.1.4, we will consider the special case $\Omega \subset \mathbb{R}^2$, $s < 1$ and $p = 2$. For this setting the norm of $u \in H^s(\Omega)$ reduces to

$$\|u\|^2_{H^s(\Omega)} = \|u\|^2_{W^{s,2}(\Omega)} = \int_\Omega \int_\Omega \frac{|u(\boldsymbol{x}) - u(\boldsymbol{y})|^2}{|\boldsymbol{x} - \boldsymbol{y}|^{2s+2}} d\boldsymbol{x} d\boldsymbol{y}.$$

2.2.3 Fourier basis

In the following chapters we want to approximate solutions to boundary value problems with truncated *Fourier series*. Here we will present some basic facts and definitions about Fourier series. Throughout all the chapters our periodic box Ω is defined as

$$\Omega := \left(-\frac{1}{2}, \frac{1}{2}\right)^d,$$

for a given dimension $d \in \mathbb{N}$. We choose an orthonormal basis of $L^2(\Omega)$, namely the trigonometric basis functions

$$\phi_{\boldsymbol{n}}(\boldsymbol{x}) := e^{\mathrm{i}2\pi \boldsymbol{n} \cdot \boldsymbol{x}}, \qquad \boldsymbol{x} \in \mathbb{R}^d,\, \boldsymbol{n} \in \mathbb{Z}^d. \tag{2.6}$$

For a function $f \in L^2(\Omega)$ the Fourier series is defined as

$$\sum_{\boldsymbol{n} \in \mathbb{Z}^d} \widehat{f}_{\boldsymbol{n}} \phi_{\boldsymbol{n}}(\boldsymbol{x}), \tag{2.7}$$

where the *Fourier coefficients* are defined as

$$\widehat{f}_{\boldsymbol{n}} := (f, \phi_{\boldsymbol{n}}) = \int_\Omega f(\boldsymbol{x})\overline{\phi_{\boldsymbol{n}}(\boldsymbol{x})}d\boldsymbol{x}. \tag{2.8}$$

According to the *Parseval identity* the corresponding L^2-norm is given by

$$\|f\|^2_{L^2(\Omega)} = \sum_{\boldsymbol{n}\in\mathbb{Z}^d} |\widehat{f}_{\boldsymbol{n}}|^2. \tag{2.9}$$

Any $\boldsymbol{u} \in \boldsymbol{L}^2(\Omega)$ can be expanded into its Fourier series

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\boldsymbol{u}}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}$$

with $\widehat{\boldsymbol{u}}_{\boldsymbol{n}} \in \mathbb{C}^3$.

2.2.4 Periodic Sobolev spaces

Since all our considerations will be in a periodic setting we need to define *periodic Sobolev spaces*. For $s \in \mathbb{R}_{\geq 0}$ we define the norm

$$\|u\|^2_{H^s_{\mathrm{per}}} := \sum_{\boldsymbol{n}\in\mathbb{Z}^d} |\boldsymbol{n}|^{2s}_\star |\widehat{u}_{\boldsymbol{n}}|^2 \quad \text{with} \quad |\boldsymbol{n}|_\star = \begin{cases} 1, & \text{if } \boldsymbol{n} = \boldsymbol{0}, \\ |\boldsymbol{n}|, & \text{if } \boldsymbol{n} \neq \boldsymbol{0}. \end{cases} \tag{2.10}$$

and the corresponding periodic Sobolev space

$$H^s_{\mathrm{per}}(\Omega) := \{u \in L^2(\Omega) : \|u\|_{H^s_{\mathrm{per}}} < \infty\}. \tag{2.11}$$

Let us discuss which condition on the decay of the Fourier coefficients $\widehat{u}_{\boldsymbol{n}}$ a function $u \in H^s_{\mathrm{per}}(\Omega)$ fulfills. Let us consider the case $d = 1$. If $u \in H^s_{\mathrm{per}}(\Omega)$ then we have

$$\sum_{n\in\mathbb{Z}} |n|^{2s}_\star |\widehat{u}_n|^2 < \infty.$$

Let

$$|\widehat{u}_n| \simeq |n|^{-t}$$

for $n \in \mathbb{Z}\backslash\{0\}$ and some $t > 0$. Then we obtain that

$$\infty > \sum_{n\in\mathbb{N}} |n|^{2s} \left(|\widehat{u}_{-n}|^2 + |\widehat{u}_n|^2\right) \simeq \sum_{n\in\mathbb{N}} n^{2s} n^{-2t} = \sum_{n\in\mathbb{N}} n^{2(s-t)}$$

needs to be satisfied. It is well known that the series

$$\sum_{n\in\mathbb{N}} n^{-\nu}$$

is convergent if and only if $\nu > 1$ is satisfied. Therefore, we obtain that

$$2(s-t) < -1 \quad \Longleftrightarrow \quad s - t < -\frac{1}{2} \quad \Longleftrightarrow \quad t > s + \frac{1}{2}$$

needs to be satisfied. This means if $u \in H^s_{\text{per}}(\Omega)$, such that for its coefficients holds

$$|\widehat{u}_n| \simeq |n|^{-t},$$

then $t > s + \frac{1}{2}$. For the two- and three-dimensional case (i.e $d = 2, 3$), with similar arguments we obtain for $u \in H^s_{\text{per}}(\Omega)$ with

$$|\widehat{u}_{\boldsymbol{n}}| \simeq |\boldsymbol{n}|^{-t},$$

that the condition $t > s + \frac{d}{2}$ needs to be satisfied. If $u \in H^s_{\text{per}}(\Omega)$, then the error made by considering a truncated Fourier series u_N instead of u can be estimated in $L^2(\Omega)$ by

$$\begin{aligned}
\|u - u_N\|^2_{L^2} &= \sum_{|\boldsymbol{n}|>N} |\widehat{u}_{\boldsymbol{n}}|^2 \\
&= \sum_{|\boldsymbol{n}|>N} |\boldsymbol{n}|^{-2s} |\boldsymbol{n}|^{2s} |\widehat{u}_{\boldsymbol{n}}|^2 \\
&< N^{-2s} \sum_{|\boldsymbol{n}|>N} |\boldsymbol{n}|^{2s} |\widehat{u}_{\boldsymbol{n}}|^2 \\
&\leq N^{-2s} \|u\|^2_{H^s_{\text{per}}}.
\end{aligned}$$

This leads to

$$\|u - u_N\|_{L^2} \lesssim N^{-s}.$$

For the vectorial setting we define

$$\boldsymbol{H}^s_{\text{per}}(\Omega) := \left\{ \boldsymbol{u} \in \boldsymbol{L}^2(\Omega) : \sum_{\boldsymbol{n} \in \mathbb{Z}^d} |\boldsymbol{n}|^{2s}_\star |\widehat{\boldsymbol{u}}_{\boldsymbol{n}}|^2 < \infty \right\}. \tag{2.12}$$

For our numerical approximations we will need appropriate subspaces of the periodic Sobolev spaces. Therefore, we define the trigonometric subspace

$$\mathcal{T}_N := \text{span}\{ e^{\mathrm{i}2\pi \boldsymbol{n} \cdot \boldsymbol{x}} \, : \, \boldsymbol{n} \in \mathbb{I}_N \}. \tag{2.13}$$

In the 3D vectorial setting later we will actually consider $(\mathcal{T}_N)^3$, however also for this setting we will denote the subspace as $\mathcal{T}_N$.

Chapter 3

Mathematical Modeling

3.1 Maxwell equations

In this section we introduce the Maxwell equations, which are a cornerstone of *classical electromagnetism*. Classical electromagnetism is the standard tool for describing the propagation of electromagnetic waves. As we are interested in the propagation of light waves in dielectric media this is the tool of interest for us. We will give a short introduction to this standard theory. For more detailed discussions on the topic we refer the interested reader to [24, 27, 33, 34, 55].

The *macroscopic form* of the Maxwell equations in SI units is given as

$$\begin{cases} \nabla \times \boldsymbol{H} - \partial_t \boldsymbol{D} = \boldsymbol{J} & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Ampère's law)}, \\ \nabla \cdot \boldsymbol{D} = \rho & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Gauss's law)}, \\ \nabla \times \boldsymbol{E} + \partial_t \boldsymbol{B} = \boldsymbol{0} & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Faraday's law)}, \\ \nabla \cdot \boldsymbol{B} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3 \quad \text{(Gauss's law for magnetism)}, \end{cases} \tag{3.1}$$

where the four vector fields $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{D}, \boldsymbol{B} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$, are the solutions we are looking for. All those vector fields are functions depending on the variables t and $\boldsymbol{x}$. As commonly used, $t \in \mathbb{R}$ is a time variable and $\boldsymbol{x} \in \mathbb{R}^3$ a space variable. The vector fields $\boldsymbol{E}$ and $\boldsymbol{H}$ are the *electric* and the *magnetic field*, the vector field $\boldsymbol{D}$ is the *electric displacement field*, and the vector field $\boldsymbol{B}$ is the *magnetic flux density*. The scalar field $\rho : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}$ and the vector field $\boldsymbol{J} : \mathbb{R} \times \mathbb{R}^3 \to \mathbb{R}^3$ are given quantities and are called *charge density* and *current density*, respectively. As in many other works, e.g. [10, 17, 22, 23, 24, 55], we consider only media with $\rho = 0$ and $\boldsymbol{J} = \boldsymbol{0}$. This means that we consider media in which no free charge is present and which are non-conducting. Since we will consider photonic crystals later, which are composed of dielectric materials, these assumptions make sense. In order to close the system of equations (3.1) so-called *constititive relations* need to be introduced. They describe how the wave interacts with the considered medium.

As in [10, 17, 22, 23, 24, 55] we only consider *linear media* which results in the following constititive relations

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \tag{3.2}$$

$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{3.3}$$

where ε is the *electric permittivity* and μ is the *magnetic permeability*, respectively. Moreover we assume that the considered medium is *lossless*, *non-dispersive* and *isotropic*. These assumptions fit to the materials we consider in later chapters. From these assumptions it follows that $\varepsilon : \mathbb{R}^3 \to \mathbb{R}_{>0}$. As the materials that we consider will be non-magnetic we assume $\mu = 1$. With these assumptions the equations (3.1) turn into

$$\begin{cases} \nabla \times \boldsymbol{H} - \varepsilon \partial_t \boldsymbol{E} = \boldsymbol{0} & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \cdot (\varepsilon \boldsymbol{E}) = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \times \boldsymbol{E} + \partial_t \boldsymbol{H} = \boldsymbol{0} & \text{in } \mathbb{R} \times \mathbb{R}^3, \\ \nabla \cdot \boldsymbol{H} = 0 & \text{in } \mathbb{R} \times \mathbb{R}^3. \end{cases} \tag{3.4}$$

3.2 Time-harmonic Maxwell equations

In the case of a monochromatic wave we can represent all the fields as a product of a function depending on the spatial variable $\boldsymbol{x} \in \mathbb{R}^3$ and a plane wave with temporal dependence. For the E-field we can write for example $\boldsymbol{E}(\boldsymbol{x})e^{\mathrm{i}\omega t}$, and for the other fields similarly. With this ansatz the equations (3.4) turn into the so-called *time-harmonic Maxwell equations*

$$\begin{cases} \nabla \times \boldsymbol{H} - \mathrm{i}\omega\varepsilon \boldsymbol{E} = \boldsymbol{0} & \text{in } \mathbb{R}^3, \\ \nabla \cdot (\varepsilon \boldsymbol{E}) = 0 & \text{in } \mathbb{R}^3, \\ \nabla \times \boldsymbol{E} + \mathrm{i}\omega \boldsymbol{H} = \boldsymbol{0} & \text{in } \mathbb{R}^3, \\ \nabla \cdot \boldsymbol{H} = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.5}$$

By eliminating one of the fields one can actually decouple these equations. If we want to eliminate $\boldsymbol{E}$ we can use from (3.5) the identity

$$\boldsymbol{E} = -\frac{\mathrm{i}}{\omega\varepsilon} \nabla \times \boldsymbol{H}.$$

Plugging in this identity into the third equation of (3.5), together with the fourth equation, this leads to the H-field formulation of the time-harmonic Maxwell equations

$$\begin{cases} \nabla \times \left(\frac{1}{\varepsilon} \nabla \times \boldsymbol{H} \right) = \omega^2 \boldsymbol{H} & \text{in } \mathbb{R}^3, \\ \nabla \cdot \boldsymbol{H} = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.6}$$

In a similar way one can eliminate $\boldsymbol{H}$ which yields the E-field formulation of the time-harmonic Maxwell equations

$$\begin{cases} \nabla \times \nabla \times \boldsymbol{E} = \omega^2 \varepsilon \boldsymbol{E} & \text{in } \mathbb{R}^3, \\ \nabla \cdot (\varepsilon \boldsymbol{E}) = 0 & \text{in } \mathbb{R}^3. \end{cases} \tag{3.7}$$

So we end up with two systems of equations, namely (3.6) and (3.7), which are constrained eigenvalues problems. These equations can be interpreted in the following way: If for any *frequency* ω no solution to the eigenvalue problems exists, a wave is not able to propagate inside the considered medium at this frequency. Notice that via the first and the third equation in (3.5) a solution of one of the problems determines the solution of the other problem. This means that we only need to solve one of the problems numerically. As our goal is to study photonic crystals, which are periodic arrangements of dielectric materials, in the following section we consider standard tools from solid state physics. These tools are useful for describing periodic structures. Similar introductions into those standard tools can be found in [10, 24, 34, 55].

3.3 Periodic structures

As we want to study wave propagation in periodic structures, namely photonic crystals, we need appropriate tools for their description. We will assume that our periodic arrangements are in whole space $\mathbb{R}^d$. This assumption is justified by the fact that the size of the periodicity cell is very small compared to the size of the whole photonic crystal. We describe periodic structures by a *Bravais lattice* Γ which is spanned by linearly independent vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_d \in \mathbb{R}^d$:

$$\Gamma := \left\{ \sum_{j=1}^{d} n_j \boldsymbol{a}_j : n_j \in \mathbb{Z}, j = 1, ..., d \right\}. \tag{3.8}$$

In our considerations $\boldsymbol{a}_i$ will always be the i-th euclidian unit vector and a photonic crystal will always be an arrangement of two different dielectric materials. We model the optical densities of the two crystal materials by two *electric permittivities*

$$\varepsilon_1, \varepsilon_2 \in \mathbb{R}, \quad 0 < \varepsilon_1 < \varepsilon_2.$$

Then we can describe the optical density of the photonic crystal by a two-valued, Γ-periodic *permittivity function*

$$\varepsilon : \mathbb{R}^3 \to \{\varepsilon_1, \varepsilon_2\}, \quad \varepsilon(\boldsymbol{x} + \boldsymbol{R}) = \varepsilon(\boldsymbol{x})$$

for all $\boldsymbol{x} \in \mathbb{R}^3$, $\boldsymbol{R} \in \Gamma$. A region in space will be called *fundamental cell* if it fills the whole space $\mathbb{R}^d$ under translations with respect to Γ. As the choice of a fundamental cell is not unique the most common choice is the *Wigner-Seitz cell* W which is defined as the region of space that is closer to the origin than to any other lattice point. Since we consider lattices $\Gamma = \mathbb{Z}^d$, depending on the dimension $d \in \mathbb{N}$, the Wigner-Seitz cell will be

$$W := \left\{ \sum_{j=1}^{d} s_j \boldsymbol{a}_j : s_j \in \left[-\frac{1}{2}, \frac{1}{2} \right], j = 1, ..., d \right\} = \left[-\frac{1}{2}, \frac{1}{2} \right]^d. \tag{3.9}$$

In this concept there exists a so-called *dual lattice*. For the sake of motivation for the choice of this dual lattice we consider the permittivity function ε that models the photonic crystal. We choose a dual lattice Γ^* such that the expansion

$$\varepsilon(\boldsymbol{x}) = \sum_{\boldsymbol{G} \in \Gamma^*} \varepsilon_{\boldsymbol{G}} e^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{x}}$$

preserves the periodicity of the crystal

$$\varepsilon(\boldsymbol{x} + \boldsymbol{R}) = \sum_{\boldsymbol{G} \in \Gamma^*} \varepsilon_{\boldsymbol{G}} e^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{x}} e^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{R}} = \varepsilon(\boldsymbol{x}),$$

with respect to Γ. This means that $\boldsymbol{G} \cdot \boldsymbol{R} \in 2\pi\mathbb{Z}$ needs to hold for every $\boldsymbol{G} \in \Gamma^*$ and $\boldsymbol{R} \in \Gamma$. Therefore, for each lattice type in the real space we define the corresponding *reciprocal lattice* Γ^* as

$$\Gamma^* := \left\{ 2\pi \sum_{j=1}^{d} n_j \boldsymbol{b}_j : n_j \in \mathbb{Z}, j = 1, ..., d \right\}, \tag{3.10}$$

with unit vectors $\boldsymbol{b}_1, ..., \boldsymbol{b}_d$ such that

$$\boldsymbol{b}_l \cdot \boldsymbol{a}_m = \delta_{lm}$$

for $l, m \in \{1, ..., d\}$. The Wigner-Seitz cell of the reciprocal lattice is called *Brillouin zone*. In our considerations, depending on the dimension $d \in \mathbb{N}$, the Brillouin zone corresponding to the reciprocal lattice Γ^* is

$$B := \left\{ \sum_{j=1}^{d} s_j \boldsymbol{b}_j : s_j \in [-\pi, \pi], j = 1, ..., d \right\} = [-\pi, \pi]^d. \tag{3.11}$$

3.4 Bloch modes

In our work we want to compute *Bloch modes* as in numerous other works, e.g. [10, 17, 22, 23, 24, 55]. For the H-field formulation (3.6) this means we want to compute eigenfunctions $\boldsymbol{H}$ of the form

$$\boldsymbol{H}(\boldsymbol{x}) = e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}), \tag{3.12}$$

where $\boldsymbol{k} \in B$ and $\boldsymbol{h}$ is Γ-periodic. With this so-called Bloch ansatz the problem (3.6) on whole $\mathbb{R}^d$ is being transformed to a parametrized eigenvalue problem on

$$\Omega := W = \left[-\frac{1}{2}, \frac{1}{2}\right]^d.$$

We set $\lambda := \omega^2$. Then the family of constrained eigenproblems reads

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k})\times \left(\dfrac{1}{\varepsilon}(\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{h}\right) = \lambda \boldsymbol{h} & \text{in } \Omega, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{h} = 0 & \text{in } \Omega, \end{cases} \tag{3.13}$$

for every $\boldsymbol{k} \in B$. Applying the Bloch ansatz

$$\boldsymbol{E}(\boldsymbol{x}) = e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\boldsymbol{e}(\boldsymbol{x}), \tag{3.14}$$

for the E-field formulation (3.7) the problem on whole of $\mathbb{R}^d$ is being transformed into a parametrized eigenvalue problem on Ω. The family of constrained eigenproblems reads

$$\begin{cases} (\nabla + \mathrm{i}\boldsymbol{k})\times (\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{e} = \lambda\varepsilon\boldsymbol{e} & \text{in } \Omega, \\ (\nabla + \mathrm{i}\boldsymbol{k}) \cdot (\varepsilon\boldsymbol{e}) = 0 & \text{in } \Omega, \end{cases} \tag{3.15}$$

for every $\boldsymbol{k} \in B$. Those two parametrized problems with the $\boldsymbol{k}$-shifted operators will be considered in our numerical computations in Chapter 7.

3.5 Photonic band structures

In this section we want to collect some results for the type of problems introduced in the previous section. The interested reader can find a more detailed discussion of the following results in [24, 34, 37, 38, 55]. For the type of periodic structures represented by a permittivity function ε, as introduced previously, one can show that for every vector $\boldsymbol{k} \in B$ the spectrum of the eigenvalue problems

$$(\nabla + \mathrm{i}\boldsymbol{k})\times \left(\frac{1}{\varepsilon}(\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{u}\right) = \lambda \boldsymbol{u} \quad \text{in } \Omega$$

or

$$(\nabla + \mathrm{i}\boldsymbol{k})\times(\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{u} = \lambda\varepsilon\boldsymbol{u} \quad \text{in } \Omega$$

is discrete with real non-negative eigenvalues $\lambda_{\boldsymbol{k},n}$ and eigenfunctions $\boldsymbol{u}_{\boldsymbol{k},n}$ for $n \in \mathbb{N}$. A detailed discussion of this topic can be found in Chapter 2 of [24], where Theorem 2.1.7 on page 27 is the statement about Maxwell eigenvalues. One can show that the curves $\boldsymbol{k} \mapsto \lambda_{\boldsymbol{k},n}$ are continuous mappings from B to $\mathbb{R}$. This holds for every ε as introduced in the section on periodic structures. The *photonic band structure* is the collection of all the curves $\boldsymbol{k} \mapsto \lambda_{\boldsymbol{k},n}$ and the spectrum of the $\boldsymbol{k}$-shifted differential operators above is

$$\sigma = \bigcup_{n\in\mathbb{N}} \left[\inf_{\boldsymbol{k}\in B} \lambda_{\boldsymbol{k},n}, \sup_{\boldsymbol{k}\in B} \lambda_{\boldsymbol{k},n}\right].$$

The general treatment of these theoretical results can be found in Chapter 3 of [24]. In the case that

$$\inf_{\boldsymbol{k}\in B} \lambda_{\boldsymbol{k},n+1} - \sup_{\boldsymbol{k}\in B} \lambda_{\boldsymbol{k},n} > 0$$

holds for some $n \in \mathbb{N}$, we say that there exists a *band gap*. These so-called band gaps are of interest. The reason for this is that for those frequencies lying in the gap a wave is not able to propagate inside the medium. If a medium has this property this can be used to create waveguides which are able to guide light waves around sharp corners.

Chapter 4

Fourier-Galerkin Method

In this chapter we want to give an introduction to the discretization of elliptic boundary value problems with the *Fourier-Galerkin method*. A similar introduction can be found in Chapter 2 of [11]. We will see that the matrices that arise from the discretization have a Toeplitz structure which can be used for efficient application of iterative methods. Therefore, after having introduced the Fourier-Galerkin discretization for elliptic boundary value problems, we will discuss how to perform matrix-vector products efficiently for the matrix structures that arise.

4.1 Galerkin discretization of elliptic bvp's

We consider the elliptic boundary value problem

$$\begin{cases} Lu = -\nabla \cdot (\nu \nabla u) + \sigma u = f & \text{in } \Omega, \\ u \text{ is } \mathbb{Z}^d\text{-periodic}, \end{cases} \tag{4.1}$$

with real coefficients ν and σ which are sufficiently smooth and σ satisfies $0 < \nu_{\min} \leq \nu(\boldsymbol{x}) \leq \nu_{\max} < \infty$ in Ω. This type of equation will play a role in Chapter 6. Variationally formulated the problem (4.1) reads: Find $u \in H^1_{\text{per}}(\Omega)$ such that

$$a(u, v) = (f, v) \qquad \forall v \in H^1_{\text{per}}(\Omega), \tag{4.2}$$

where

$$a(u, v) := \int_\Omega \nu \nabla u \cdot \nabla \overline{v} d\boldsymbol{x} + \int_\Omega \sigma u \overline{v} d\boldsymbol{x}. \tag{4.3}$$

Choosing $N \in \mathbb{N}$, and thus the finite dimensional subspace $\mathcal{T}_N$ of $H^1_{\text{per}}(\Omega)$, the Galerkin approximation is defined in the following way: Find $u_N \in \mathcal{T}_N$ such that

$$a(u_N, v_N) = (f, v_N) \qquad \forall v_N \in \mathcal{T}_N. \tag{4.4}$$

In order to transform the problem (4.4) into an algebraic problem we use the representation

$$u_N(\boldsymbol{x}) = \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}}\phi_{\boldsymbol{n}}(\boldsymbol{x}) = \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}$$

and then test this equation with all the basis functions of $\mathcal{T}_N$. Due to linearity this means we have to consider

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} a\big(\phi_{\boldsymbol{n}}, \phi_{\boldsymbol{m}}\big) = b\big(f, \phi_{\boldsymbol{m}}\big)$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. With the definition of the Fourier coefficients we obtain

$$\begin{aligned} a\big(\phi_{\boldsymbol{n}}, \phi_{\boldsymbol{m}}\big) &= 4\pi^2 \boldsymbol{n}\cdot\boldsymbol{m} \int_\Omega \nu(\boldsymbol{x}) e^{-\mathrm{i}2\pi(\boldsymbol{m}-\boldsymbol{n})\cdot\boldsymbol{x}} d\boldsymbol{x} + \int_\Omega \sigma(\boldsymbol{x}) e^{-\mathrm{i}2\pi(\boldsymbol{m}-\boldsymbol{n})\cdot\boldsymbol{x}} d\boldsymbol{x} \\ &= 4\pi^2 \boldsymbol{n}\cdot\boldsymbol{m}\, \widehat{\nu}_{\boldsymbol{m}-\boldsymbol{n}} + \widehat{\sigma}_{\boldsymbol{m}-\boldsymbol{n}} \end{aligned}$$

and

$$(f, \phi_{\boldsymbol{m}}) = \int_\Omega f(\boldsymbol{x}) e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}} d\boldsymbol{x} = \widehat{f}_{\boldsymbol{m}}.$$

So the problem (4.4) transforms to

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} 4\pi^2 \boldsymbol{n}\cdot\boldsymbol{m}\, \widehat{\nu}_{\boldsymbol{m}-\boldsymbol{n}} \widehat{u}_{\boldsymbol{n}} + \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\sigma}_{\boldsymbol{m}-\boldsymbol{n}} \widehat{u}_{\boldsymbol{n}} = \widehat{f}_{\boldsymbol{m}} \tag{4.5}$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. If the coefficients ν and σ are not constant, then this leads to such convolutional sums as in (4.5). The second sum in (4.5) can be interpreted as the $\boldsymbol{m}$-th component of a matrix-vector product $S\widehat{\boldsymbol{u}}$, where the matrix S is a BTTB matrix containing Fourier coefficients of σ and the vector $\widehat{\boldsymbol{u}}$ contains the Fourier coefficients of u. In the first sum of (4.5) we also have such a convolutional sum, however with additional terms. This can be represented as a product of block-diagonal matrices and a BTTB matrix. As performing a matrix-vector product with a sparse matrix does not cause difficulties we skip the discussion of this issue to Section 6.1.3. In the remaining sections of this chapter we will discuss how matrix-vector products with Toeplitz and BTTB matrices can be performed efficiently.

4.2 Fast Toeplitz multiplication

In general performing a matrix-vector product $A\boldsymbol{x}$ for $A \in \mathbb{R}^{N\times N}$ and $\boldsymbol{x} \in \mathbb{R}^N$ has the computational cost of $\mathcal{O}(N^2)$ operations. However, if a matrix has a Toeplitz structure, as we will have later in our discretizations, then this cost can be reduced

to $\mathcal{O}(N\log(N))$ operations. We want to discuss the concepts in this section. These are well known concepts and can be found similarly in [26, 69]. For a thorough discussion of this topic we refer to Section 4.2 in [69]. For the discussions about BTTB matrices we will follow the presentation in the appendix of [26], where central results for circulant matrices from [19] were used.

4.2.1 Fast convolution

In this subsection we will introduce some concepts that will enable us to compute a matrix-vector product in $\mathcal{O}(N\log(N))$ operations, if a certain structure of the matrix is given. The following tools will play a central role in those fast matrix-vector products.

Definition 4.2.1. *Let $\omega_N := e^{-\mathrm{i}2\pi/N}$. The isomorphism*

$$\mathcal{F}_N : \mathbb{C}^N \to \mathbb{C}^N, \quad (f_j) \mapsto (\widehat{f_j})$$

with

$$\widehat{f_k} = \sum_{j=0}^{N-1} f_j e^{-2\pi \mathrm{i} jk/N} = \sum_{j=0}^{N-1} f_j \omega_N^{kj} \quad \text{for } k = 0, ..., N-1$$

*is called **discrete Fourier transform** (DFT). The inverse mapping $\mathcal{F}_N^{-1}$ is*

$$f_j = \frac{1}{N} \sum_{k=0}^{N-1} \widehat{f_k} e^{2\pi \mathrm{i} jk/N} \quad \text{for } j = 0, ..., N-1$$

*and is called **inverse discrete Fourier transform** (IDFT). The matrix $F_N := (\omega_N^{kj})_{kj}$ corresponding to $\mathcal{F}_N$ is called **Fourier matrix**. The linear mapping $\mathcal{F}_N : \mathbb{C}^N \to \mathbb{C}^N$ can be written as matrix-vector multiplication:*

$$\widehat{\boldsymbol{f}} = F_N \boldsymbol{f}.$$

So the discrete Fourier transform of a vector $\boldsymbol{f}$ can be interpreted as a matrix-vector product with the matrix F_N, which takes $\mathcal{O}(N^2)$ operations. The matrix F_N has the properties $F_N^H F_N = N I_N$ and thus $F_N^{-1} = \frac{1}{N} F_N^H$, see for example Section 1.1 in [69], Section 6.7.1. in [20] or Section 2.3.1 in [65]. A well-known tool is the famous *fast Fourier transform* (FFT) by Cooley and Tukey [15]. The presentation of this algorithm was a huge breakthrough in computational mathematics in the last century, which is used in many technical applications nowadays. If $N = 2^l$, for some $l \in \mathbb{N}$, it actually computes the same as the discrete Fourier transform, however, more efficiently, namely in $\mathcal{O}(N\log(N))$ operations. For a detailed discussion of the FFT and many aspects connected to it we refer to van Loan's standard textbook [69] on this subject. Next we define an operation, which can be interpreted as a matrix-vector product, where the matrix has a special structure.

Definition 4.2.2. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$, where $\boldsymbol{x}$ is periodically extended. The cyclic convolution $\boldsymbol{x} * \boldsymbol{y} \in \mathbb{C}^N$ is defined as*

$$(\boldsymbol{x} * \boldsymbol{y})_k := \sum_{j=0}^{N-1} x_{k-j} y_j \qquad \text{for } k = 1, ..., N.$$

The cyclic convolution of two vectors can be interpreted as a product of a circulant matrix with a vector, namely

$$\boldsymbol{x} * \boldsymbol{y} = \begin{pmatrix} x_0 & x_{N-1} & \dots & x_2 & x_1 \\ x_1 & x_0 & \ddots & x_3 & x_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ x_{N-2} & x_{N-3} & \ddots & x_0 & x_{N-1} \\ x_{N-1} & x_{N-2} & \dots & x_1 & x_0 \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-2} \\ y_{N-1} \end{pmatrix}. \tag{4.6}$$

The computational cost of a cyclic convolution is $\mathcal{O}(N^2)$ operations. For the definition of a circulant matrix see (2.5). A very important result on our way to a more efficient Toeplitz multiplication is the so-called *convolution theorem*:

Theorem 4.2.3. *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$, where $\boldsymbol{x}$ is periodically extended, it holds*

$$\mathcal{F}_N(\boldsymbol{x} * \boldsymbol{y}) = \mathcal{F}_N \boldsymbol{x} \odot \mathcal{F}_N \boldsymbol{y}.$$

Proof. With the definition of the discrete Fourier transform and the convolution we obtain

$$\begin{aligned} (\mathcal{F}_N(\boldsymbol{x} * \boldsymbol{y}))_m &= \sum_{k=0}^{N-1} \omega^{mk} \sum_{j=0}^{N-1} x_{k-j} y_j \\ &\overset{l=k-j}{=} \sum_{j=0}^{N-1} \sum_{l=-j}^{N-1-j} \omega^{m(l+j)} x_l y_j \\ &= \sum_{j=0}^{N-1} \omega^{mj} y_j \cdot \sum_{l=0}^{N-1} \omega^{ml} x_l \\ &= \widehat{\boldsymbol{y}}_m \widehat{\boldsymbol{x}}_m, \end{aligned}$$

where $\widehat{\boldsymbol{x}} = \mathcal{F}_N \boldsymbol{x}$ and $\widehat{\boldsymbol{y}} = \mathcal{F}_N \boldsymbol{y}$. □

Corollary 4.2.4. *For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^N$, where $\boldsymbol{x}$ is periodically extended, it holds*

$$\boldsymbol{x} * \boldsymbol{y} = \mathcal{F}_N^{-1}(\mathcal{F}_N \boldsymbol{x} \odot \mathcal{F}_N \boldsymbol{y}).$$

These results tell us that a cyclic convolution actually can be done in $\mathcal{O}(N\log(N))$ instead of $\mathcal{O}(N^2)$ operations. This can be realized by computing the FFTs $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$ of the vecors $\boldsymbol{x}$ and $\boldsymbol{y}$, and then by computing the inverse FFT of the pointwise product of $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{y}}$, in order to obtain $\boldsymbol{x} * \boldsymbol{y}$. The FFTs need $\mathcal{O}(N\log(N))$, the pointwise product $\mathcal{O}(N)$, resulting in $\mathcal{O}(N\log(N))$ operations. In the next subsection we will discuss how we can use this fact in order to reduce the computational cost for a matrix-vector product with a Toeplitz matrix.

4.2.2 Extension of Toeplitz to circulant matrix

We have seen that circulant matrices have desirable properties. Next we show how any Toeplitz matrix

$$T = \begin{pmatrix} t_0 & t_{-1} & t_{-2} & \dots & \dots & t_{1-n} \\ t_1 & t_0 & t_{-1} & \ddots & & t_{2-n} \\ t_2 & t_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & t_{-1} & t_{-2} \\ \vdots & & \ddots & t_1 & t_0 & t_{-1} \\ t_{n-1} & \dots & \dots & t_2 & t_1 & t_0 \end{pmatrix}$$

of order $n \times n$ can be embedded into a circulant matrix $C \in \mathbb{C}^{N\times N}$, where N is the smallest power of 2 (for the FFT) such that $N \geq 2n-1$. Let $C \in \mathbb{C}^{N\times N}$ be the circulant matrix whose first column is given by

$$(t_0, ..., t_{n-1}, 0, ..., 0, t_{1-n}, ..., t_{-1})^\top.$$

The number of zeros is $N-(2n-1)$. Then C is a circulant matrix which contains T as its upper left $N \times N$ block. If for a given $\boldsymbol{x} \in \mathbb{R}^n$ we are interested in $T\boldsymbol{x}$ we can compute

$$C\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} T\boldsymbol{x} \\ * \end{pmatrix}$$

in order to obtain $T\boldsymbol{x}$ in $\mathcal{O}(n\log(n))$ operations instead of $\mathcal{O}(n^2)$. Since we never work with matrices but rather with FFTs of vectors of roughly double the length this is an acceptable price that we pay for the reduction of $\mathcal{O}(n^2)$ to $\mathcal{O}(n\log(n))$ operations.

4.2.3 Block Toeplitz matrices

The concept of fast Toeplitz multiplication can be extended to BTTB matrices. If we think of the Fourier-Galerkin method introduced in Section 4.1, then in 1D a

Toeplitz matrix occurs in the discretization due to the convolution of the Fourier coefficients. In 2D this convolution turns into a matrix-vector product with a BTTB. In 3D the blocks of the BTTB matrix are themselves BTTB matrices. Here we explain how to generalize the concept from 1D to 2D. The extension from 2D to 3D can be done in the same fashion. Since we only work with quadratic matrices we will only treat this case, however the other case can be done similarly. An $N^2 \times N^2$ BTTB is of the form

$$T = \begin{pmatrix} T_0 & T_{-1} & T_{-2} & \dots & \dots & T_{1-N} \\ T_1 & T_0 & T_{-1} & \ddots & & T_{2-N} \\ T_2 & T_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & T_{-1} & T_{-2} \\ \vdots & & \ddots & T_1 & T_0 & T_{-1} \\ T_{N-1} & \dots & \dots & T_2 & T_1 & T_0 \end{pmatrix}, \tag{4.7}$$

where each T_j, $j = 1-N, ..., N-1$, is a Toeplitz matrix as defined in (2.3).

4.2.4 Fast multiplication for block Toeplitz matrices

We want to extend the concept for usual Toeplitz matrices to BTTB matrices. In order to reduce the computational cost for a matrix-vector product we will use the FFT again. As we are now in a two-dimensional convolution case we define analogously as for the 1D case the 2D DFT. If $f \in \mathbb{C}^{M \times N}$ then the DFT of this two-dimensional array is defined as

$$\widehat{f}_{jk} := \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f_{mn} e^{-\mathrm{i}2\pi jm/M} e^{-\mathrm{i}2\pi kn/N} \tag{4.8}$$

for $j = 0, ..., M-1$ and $k = 0, ..., N-1$. The inverse DFT is given as

$$f_{mn} = \frac{1}{M}\frac{1}{N} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} \widehat{f}_{kl} e^{\mathrm{i}2\pi mk/M} e^{\mathrm{i}2\pi nl/N} \tag{4.9}$$

for $m = 0, ..., M-1$ and $n = 0, ..., N-1$. Analogously one can make the same definitions in higher dimensions. As for the 1D DFT there exists a FFT for the 2D and 3D case, which we shall work with. It is built in as standard tool into MATLAB®, which we will permanently use in our computations. Using the Kronecker product we can write (check Section 3.4 in [69] or the appendix of [26]) the DFT of a two-dimensional array f as

$$\mathbf{vec}\left(\widehat{f}\right) = (F_M \otimes F_N)\mathbf{vec}(f)$$

and thus the IDFT as

$$\mathbf{vec}(f) = (F_M \otimes F_N)^{-1}\mathbf{vec}\left(\widehat{f}\right).$$

Now we consider the generalization of circulant matrices to BCCB matrices. An $MN \times MN$ block matrix of the form

$$C = \begin{pmatrix} C_0 & C_{M-1} & \dots & C_2 & C_1 \\ C_1 & C_0 & \ddots & C_3 & C_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ C_{M-2} & C_{M-3} & \ddots & C_0 & C_{M-1} \\ C_{M-1} & C_{M-2} & \dots & C_1 & C_0 \end{pmatrix}, \tag{4.10}$$

where the blocks C_j, $j = 0, ..., M-1$, are circulant $N \times N$ matrices is called a *block circulant matrix with circulant blocks* (BCCB). As for the case of a usual circulant matrix, a BCCB is completely determined by its first column. Let B be the matrix whose columns are the first columns of the circulant blocks $C_0, ..., C_{M-1}$. Then the 2D-DFT $\widehat{B}$ of B can be written as

$$\mathbf{vec}\left(\widehat{B}\right) = (F_M \otimes F_N)\mathbf{vec}(B).$$

By Theorem 5.8.1 in [19] the BCCB matrix C has has the diagonalization

$$C = (F_M \otimes F_N)^{-1} D (F_M \otimes F_N),$$

where

$$D = \operatorname{diag}\left(\mathbf{vec}\left(\widehat{B}\right)\right).$$

We can use this to rewrite the matrix-vector product of C with a vector $\boldsymbol{x}$ of length MN in the following way:

$$C\boldsymbol{x} = (F_M \otimes F_N)^{-1} D (F_M \otimes F_N)\boldsymbol{x} = (F_M \otimes F_N)^{-1}\left(\mathbf{vec}\left(\widehat{B}\right) \odot (F_M \otimes F_N)\boldsymbol{x}\right).$$

In this representation of the matrix-vector product we can see the same procedure as for usual circulant matrices. The product can be computed by applying two 2D-DFTs, a pointwise product and finally an inverse 2D-DFT. Altogether we can compute this product in $\mathcal{O}(MN\log(MN))$ operations via the FFT. If we have a BTTB matrix, and want to compute matrix-vector products with the FFT efficiently, we need to embed our BTTB matrix into a BCCB matrix to make the results above applicable. Now consider a $mn \times mn$ BTTB matrix T, where each Toeplitz block is $n \times n$. We can embed each Toeplitz block into a circulant block, where N is the smallest power of 2 satisfying $N \geq 2n - 1$. Now the embedding

of the BTTB matrix T can be done analogously to what we did for usual Toeplitz matrices in Section 4.2.2. The only difference is that we have to work with blocks instead of scalars. This means we end up with an $MN \times MN$ BCCB matrix, where M is the smallest power of 2 such that $M \geq 2m - 1$ holds. If we now want to compute $T\boldsymbol{x}$, then we also have to embed $\boldsymbol{x}$ into a larger vector appropriately. We consider $\boldsymbol{x}$ to be a vector consisiting of m blocks of length n stacked one below another. Now we have to append $N - n$ zeros after each block, and after that an additional zero vector of length $(M - m)N$. This way one can compute $T\boldsymbol{x}$ in $\mathcal{O}(mn \log(mn))$ operations instead of $\mathcal{O}(m^2 n^2)$ operations.

Chapter 5

Fourier factorization

In this chapter we discuss Li's factorization rules which were introduced in [41]. Li showed that when working with discontinuous structures and Fourier series, convergence issues do not only arise from the fact that Fourier series are being used. He rather showed that appropriate factorization rules need to be applied when Fourier coefficients are convolved. We will discuss the theorems with all the proofs, which can be found as a shorter version in the appendix of Chapter 4 of [7]. Most of this chapter is adopted from Chapter 4 of [7], however in much more detail. We find that these theorems are a very nice example how numerical simulation of physical phenomena can lead to theoretical results in pure mathematics. In order to keep this chapter self-contained we will introduce all definitions from [41] and [7]. Moreover we will also introduce the necessary tools for the proofs. The reason why we treat this result by Li in great detail is that we hoped to be able to apply his inverse convolution rule even in situations when at first sight it does not seem desirable. In Chapter 7 we will see that it actually would be desirable in some situations to use this rule. In order to understand why this does not work it is crucial to understand the proofs of Li's Fourier factorization theorems. At the end of this chapter we will illustrate the theorems with several examples.

5.1 Notations and conventions

Before stating the theorems we want to become familiar with the definitions that are needed for the theorems.

Definition 5.1.1. *Let $\boldsymbol{P}$ be the set of $2\pi-$periodic, real valued functions which are piecewise in C^2, i.e. an $m \in \mathbb{N}$ exists and $a_k \in [0, 2\pi]$ for $k = 0, ..., m$ exist such that $0 = a_0 < a_1 < ... < a_m = 2\pi$ with $f_k \in C^2(a_k, a_{k+1})$ for $k = 0, ..., m-1$.*

Notice that if $f \in \boldsymbol{P}$, $g \in \boldsymbol{P}$ and $h(x) := f(x)g(x)$, then $h \in \boldsymbol{P}$. We will keep this notation throughout this chapter, this means everytime we talk about a function

h we always consider h to be the product of two functions f and g. The next three definitions are from Section 4.4.3 in [7]. Since we are considering functions with jump discontinuities we need the following definition for the jump locations:

Definition 5.1.2. *For $f \in \boldsymbol{P}$ let*

$$U_f := \{x_j \mid f(x_j+) \neq f(x_j-), x_j \in [0, 2\pi), j = 1, 2, ...\}$$

*be the **set of the jump locations** of f, and let U_g be similarly defined for g. Then*

$$U_{f,g} := U_f \cap U_g$$

*is the set of **concurrent discontinuities** of f and g.*

If f and g have jumps at the same location such that their product is continuous, then the following definition applies.

Definition 5.1.3. *Let $h = fg$. If h is such that*

$$h(x_j-) = h(x_j+) \qquad \text{for all } x_j \in U_{f,g},$$

*f and g are said to have a pair of **complementary jumps** at x_j.*

In order to have a short notation for the size of the jumps the following definition is useful:

Definition 5.1.4. *The size of the jump at $x_j \in U_f$ will be denoted by $f_j^{\sqcap}$, i.e.*

$$f_j^{\sqcap} := f(x_j+) - f(x_j-).$$

As we are considering complex Fourier series, we are considering series which are indexed by all integers, see e.g. [70]. In classical analysis a series

$$\sum_{n\in\mathbb{Z}} a_n$$

is considered to be covergent if both

$$\sum_{n=0}^{\infty} a_n$$

and

$$\sum_{n=1}^{\infty} a_{-n}$$

are convergent. As we are dealing with complex Fourier series we consider symmetric partial sums

$$s_M := \sum_{n=-M}^{M} a_n,$$

which means

$$\sum_{n\in\mathbb{Z}} a_n := \lim_{M\to\infty} s_M = \lim_{M\to\infty} \sum_{n=-M}^{M} a_n.$$

Considering such symmetric series makes sense, since there exist series which are divergent in the classical sense but their symmetric partial sums converge. A simple example is

$$\sum_{n=-\infty}^{\infty} n.$$

During the proof of the Fourier factorization theorems we will have to deal with symmetric sums. This means that whenever we consider sums with indices over all integers or a sum with the same amount of positive and negative indices it is important to keep in mind the definition for symmetric sums.

5.2 Convolution of Fourier series

In the discussion of [41] the Fourier series of the fields and of the permittivity are inserted into the Maxwell equations, which leads to the question how to compute the Fourier coefficients of h if the Fourier coefficients of f and g are given. This is answered by Laurent's rule [73] which roughly states that if f_n are the Fourier coefficients of f, and g_n are the Fourier coefficients of g, respectively, the Fourier coefficients h_n of h can be computed in the following way:

$$h_n = \sum_{m=-\infty}^{+\infty} f_{n-m} g_m.$$

This means

$$h(x) = \lim_{N\to\infty} \sum_{n=-N}^{+N} \left(\lim_{M\to\infty} \sum_{m=-M}^{+M} f_{n-m} g_m \right) e^{\mathrm{i}nx}.$$

Due to the fact that in practice the series have to be truncated at some index the common choice is to truncate the series symmetrically, i.e. $N = M$. This leads to the question whether

$$h(x) \stackrel{?}{=} \lim_{M\to\infty} \sum_{n=-M}^{+M} \left(\sum_{m=-M}^{+M} f_{n-m} g_m \right) e^{\mathrm{i}nx}$$

holds. The answer to this question was given in form of the Fourier Factorization Theorems by Li [41]. Before stating the theorems and their proofs we make the following definitions for the truncated series:

$$\begin{aligned} h_n^{(M)} &:= \sum_{m=-M}^{+M} f_{n-m} g_m, \\ h^{(M)}(x) &:= \sum_{n=-M}^{+M} h_n^{(M)} e^{\mathrm{i}nx}, \\ h_M(x) &:= \sum_{n=-M}^{+M} h_n e^{\mathrm{i}nx}. \end{aligned}$$

This means $h_n^{(M)}$ is the Fourier coefficient computed by the truncated Laurent rule, $h^{(M)}(x)$ is the truncated Fourier series with $h_n^{(M)}$ as Fourier coefficients and $h_M(x)$ is the truncated Fourier series with the exact Fourier coefficients.

5.3 Useful tools

As a preparation for the thereoms on Fourier factorization in this chapter we collect several useful theorems, definitions, etc. First we state Abel's theorem on partial summation, which can be found in [31].

Theorem 5.3.1 ([31], Theorem 11.2)**.** *Let $n \in \mathbb{N}$ and $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{R}$. Then the following identity holds:*

$$\sum_{k=1}^{n} a_k b_k = A_n b_n + \sum_{k=1}^{n-1} A_k (b_k - b_{k+1}),$$

with $A_k = a_1 + a_2 + \ldots + a_k$.

The next theorem, which can be found in [36], allows to check the convergence of a series via an integral test.

Theorem 5.3.2 ([36], Theorem 176)**.** *Assume that $f(x) \geq 0$ and that f decreases monotonically on $[m, \infty)$, for all $m \in \mathbb{N}$. Then*

$$\int_m^{\infty} f(x) dx$$

converges if and only if

$$\sum_{n=m}^{\infty} f(n)$$

converges. Moreover, in case of convergence, the following inequalities hold:

$$\sum_{n=m+1}^{\infty} f(n) \leq \int_m^{\infty} f(x)\mathrm{d}x \leq \sum_{n=m}^{\infty} f(n). \tag{5.1}$$

For the proof of the Fourier factorization theorems we will need to know how the Fourier coefficients of a function $f \in \boldsymbol{P}$, which is also continuous, decay. This will be answered in the following theorem.

Theorem 5.3.3. *Let $f \in \boldsymbol{P}$ and let f be continuous. Then the following inequality for the decay of the Fourier coefficients of f holds*

$$\left|\widehat{f}_n\right| \leq \frac{C_f}{n^2},$$

where C_f depends on $\|f_k'\|_\infty$ and $\|f_k''\|_{L^2}$.

Proof. Let $0 = a_0 < a_1 < ... < a_m = 2\pi$ and $f \in C^2(a_k, a_{k+1})$ for $k = 0, ..., m-1$. Then we obtain

$$\widehat{f}_n = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-\mathrm{i}nt}dt = \frac{1}{2\pi}\sum_{k=0}^{m-1}\int_{a_k}^{a_{k+1}} f_k(t)e^{-\mathrm{i}nt}dt,$$

with

$$\begin{aligned}\int_{a_k}^{a_{k+1}} f_k(t)e^{-\mathrm{i}nt}dt &= \left[\frac{\mathrm{i}}{n}f_k(t)e^{-\mathrm{i}nt}\right]_{a_k}^{a_{k+1}} - \frac{\mathrm{i}}{n}\int_{a_k}^{a_{k+1}} f_k'(t)e^{-\mathrm{i}nt}dt \\ &= \left[\frac{\mathrm{i}}{n}f_k(t)e^{-\mathrm{i}nt}\right]_{a_k}^{a_{k+1}} + \left[\frac{1}{n^2}f_k'(t)e^{-\mathrm{i}nt}\right]_{a_k}^{a_{k+1}} \\ &\quad -\frac{1}{n^2}\int_{a_k}^{a_{k+1}} f_k''(t)e^{-\mathrm{i}nt}dt.\end{aligned}$$

Since f is 2π-periodic and continuous we obtain

$$\begin{aligned}\widehat{f}_n &= \frac{1}{2\pi}\sum_{k=0}^{m-1}\left(\left[\frac{1}{n^2}f_k'(t)e^{-\mathrm{i}nt}\right]_{a_k}^{a_{k+1}} - \frac{1}{n^2}\int_{a_k}^{a_{k+1}} f_k''(t)e^{-\mathrm{i}nt}dt\right) \\ &\leq \frac{m}{2\pi n^2}\left(\max_k \|f_k'\|_\infty + \max_k \|f_k''\|_{L^2}\right),\end{aligned}$$

which leads to

$$\left|\widehat{f}_n\right| \leq \frac{C_f}{n^2}$$

with

$$C_f = \frac{m}{2\pi}\left(\max_k \|f_k'\|_\infty + \max_k \|f_k''\|_{L^2}\right).$$

□

In the next theorem we present an important result from Fourier analysis. This theorem can be found in Zygmund's standard textbook on trigonometric series [73] (page 90, Theorem 3.7). This result will be useful for us because it gives us the boundedness of a series, which we will have to estimate in the proof of the factorization theorems later.

Theorem 5.3.4 ([73], Theorem 3.7)**.** *If the 2π-periodic function f is of bounded variation then the partial Fourier sums of f, i.e.*

$$S_{f,N}(x) := \sum_{n=-N}^{N} \widehat{f}_n e^{-\mathrm{i}nx}$$

are uniformly bounded in x.

The following Lemma is a well-known result from complex analysis and will be needed for showing the boundedness of some terms several times in the proof.

Lemma 5.3.5. *Let $x \in (0, 2\pi)$ and $L, M \in \mathbb{N}$. Then*

$$\left| \sum_{k=L}^{M} e^{\mathrm{i}kx} \right| \leq C(x),$$

where $C(x)$ is a constant depending on x, but not on M or L, respectively.

Proof. Since $e^{\mathrm{i}kx} = (e^{\mathrm{i}x})^k$, we have by the geometrical series formula

$$\begin{aligned} \left| \sum_{k=L}^{M} e^{\mathrm{i}kx} \right| &= \left| \frac{e^{\mathrm{i}Lx} - e^{\mathrm{i}(M+1)x}}{1 - e^{\mathrm{i}x}} \right| \\ &= \left| \frac{e^{\mathrm{i}Lx}}{1 - e^{\mathrm{i}x}} (1 - e^{\mathrm{i}(M+1-L)x}) \right| \\ &\leq \frac{2}{|1 - e^{\mathrm{i}x}|} \\ &= \frac{\sqrt{2}}{\sqrt{1 - \cos(x)}} =: C(x), \end{aligned}$$

where the last equality follows from Euler's theorem. □

Finally we state a theorem from complex analysis which is due to Abel. This theorem allows to check the convergence of series which occur in complex analysis. Especially one estimate in the proof of the theorem will be important for the proof of the theorems in the next chapter.

Theorem 5.3.6. *Let*

$$f(z) := \sum_{n=0}^{\infty} a_n z^n,$$

with $a_n > 0$ *and* $a_n \searrow 0$ $(n \to \infty)$*. Then the power series converges if* $|z| \leq 1$ *and* $z \neq 1$*.*

Proof. We define $s_n(z) := 1 + z + \cdots + z^n$. It is easy to show that

$$s_n(z) = \frac{1 - z^{n+1}}{1 - z}$$

for $|z| \leq 1$ and $z \neq 1$. With the triangle inequality we make the following estimate

$$|s_n(z)| \leq \frac{2}{|1 - z|} =: v(z), \quad |z| \leq 1, \quad z \neq 1.$$

For $n \geq m \geq 1$ holds

$$\begin{aligned}
\sum_{k=m}^{n} a_k z^k &= \sum_{k=m}^{n} a_k \left(s_k(z) - s_{k-1}(z)\right) \\
&= \sum_{k=m}^{n} a_k s_k(z) - \sum_{k=m-1}^{n-1} a_{k+1} s_k(z) \\
&= \sum_{k=m}^{n-1} \left(a_k - a_{k+1}\right) s_k(z) + a_n s_n(z) - a_m s_{m-1}(z).
\end{aligned}$$

Since (a_n) is monotonically decreasing we obtain

$$\begin{aligned}
\left|\sum_{k=m}^{n} a_k z^k\right| &\leq \sum_{k=m}^{n-1} \left(a_k - a_{k+1}\right) v(z) + \left(a_n + a_m\right) v(z) \\
&= \left(a_m - a_n + a_n + a_m\right) v(z) \\
&= 2a_m v(z). \qquad (5.2)
\end{aligned}$$

Due to the fact that a_m tends to zero for $m \to \infty$, with Cauchy's criterion [56] we obtain the convergence of the power series. □

5.4 Fourier factorization theorems

In this section it is important to keep in mind the definitions of the first section. Moreover, remember that h is always the product of f and g, where $f, g \in \boldsymbol{P}$.

In order to keep the notation at a reasonable level for the rest of this chapter we will omit the hat for the Fourier coefficients. This means when we talk about a function f, then rather f_n will symbolize the n-th Fourier coefficient than $\widehat{f}_n$. We will state the first two theorems and prove them together in one proof. The first theorem considers the product of two discontinuous functions which do not have jumps at the same locations.

Theorem 5.4.1 ([7], Theorem 4.3)**.** *If $f \in \boldsymbol{P}$ and $g \in \boldsymbol{P}$ have no concurrent jump discontinuities and $h_n^{(M)}$ is given by*

$$h_n^{(M)} = \sum_{m=-M}^{+M} f_{n-m} g_m,$$

then

$$\lim_{M\to\infty} h^{(M)}(x) = h(x).$$

The second theorem considers the product of two discontinuous functions which have jumps at the same locations.

Theorem 5.4.2 ([7], Theorem 4.4)**.** *If $f \in \boldsymbol{P}$ and $g \in \boldsymbol{P}$ have concurrent jump discontinuities and $h_n^{(M)}$ is given by*

$$h_n^{(M)} = \sum_{m=-M}^{+M} f_{n-m} g_m,$$

then

$$h^{(M)}(x) = h_M(x) - \sum_{x_p \in U_{f,g}} \frac{f_p^{\sqcap} f_p^{\sqcap}}{2\pi^2} \Phi_M(x - x_p) - o(1),$$

where the term $o(1)$ uniformly tends to zero for $M \to \infty$, and

$$\Phi_M(z) := \sum_{n=1}^{M} \frac{\cos(nz)}{n} \sum_{|m|>M} \frac{1}{m-n}. \tag{5.3}$$

Furthermore,

$$\lim_{M\to\infty} \Phi_M(z) = \begin{cases} 0 & (z \neq 0), \\ \frac{\pi^2}{4} & (z = 0). \end{cases}$$

Proof. We begin with the idea to decompose a function $f \in \boldsymbol{P}$ into a linear part, which is discontinuous, and into a continuous part. We decompose $f \in \boldsymbol{P}$ as follows:

$$f(x) = \widetilde{f}(x) + \sum_{x_j \in U_f} \frac{f_j^{\sqcap}}{\pi} \phi(x - x_j),$$

where $\widetilde{f}$ is the continuous part and the discontinuous part is represented with the help of (periodically extended)

$$\phi(x) := \frac{1}{2}(\pi - x) \quad (0 < x < 2\pi).$$

If we define

$$\begin{aligned} Q(x) &:= \widetilde{f}(x)\widetilde{g}(x), \\ R(x;x_j) &:= \phi(x-x_j)\widetilde{g}(x), \\ S(x;x_k) &:= \widetilde{f}(x)\phi(x-x_k), \\ T(x;x_j;x_k) &:= \phi(x-x_j)\phi(x-x_k), \end{aligned}$$

the function $h = fg$ can be written as

$$\begin{aligned} h(x) &= Q(x) + \frac{1}{\pi}\sum_{x_j \in U_f} f_j^{\sqcap} R(x;x_j) + \frac{1}{\pi}\sum_{x_k \in U_g} g_k^{\sqcap} S(x;x_k) \\ &\quad + \frac{1}{\pi^2}\sum_{\substack{x_j \in U_f \\ x_k \in U_g}} f_j^{\sqcap} g_k^{\sqcap} T(x;x_j;x_k). \end{aligned}$$

The Fourier coefficients of ϕ are $\phi_0 = 0$ and $\phi_m = 1/(2\mathrm{i}m)$ for $m \neq 0$. With Theorem 5.3.3 we obtain for continuous functions $\gamma \in \boldsymbol{P}$, that for the Fourier coefficients $\gamma_m = \mathcal{O}(1/m^2)$ holds. Next, we prove several estimates which hold for $|n| \leq M$ and $0 \leq x < 2\pi$. We begin with

$$\begin{aligned} \left|Q_n^{(M)} - Q_n\right| &= \left|\sum_{m=-M}^{M} \widetilde{f}_{n-m}\widetilde{g}_m - \sum_{m=-\infty}^{\infty} \widetilde{f}_{n-m}\widetilde{g}_m\right| = \left|\sum_{|m|>M} \widetilde{f}_{n-m}\widetilde{g}_m\right| \\ &\leq \sum_{|m|>M} \left|\widetilde{f}_{n-m}\widetilde{g}_m\right| \leq \sum_{|m|>M} \left|\frac{C_f}{(n-m)^2}\frac{C_g}{m^2}\right| \\ &\leq \frac{\widetilde{C}}{M^2}\sum_{|m|>M}\frac{1}{(n-m)^2} \leq \frac{\widetilde{C}}{M^2}\frac{\pi^2}{3} \\ &\leq \frac{C}{M^2}, \end{aligned}$$

and it follows

$$\left|Q^{(M)}(x) - Q_M(x)\right| = \left|\sum_{n=-M}^{M} Q_n^{(M)} e^{\mathrm{i}nx} - \sum_{m=-M}^{M} Q_n e^{\mathrm{i}nx}\right|$$

$$\begin{aligned} &= \left| \sum_{n=-M}^{M} \left(Q_n^{(M)} - Q_n \right) e^{\mathrm{i}nx} \right| \\ &\leq \sum_{|n|\leq M} \left| Q_n^{(M)} - Q_n \right| \\ &\leq \sum_{|n|\leq M} \frac{C}{M^2} = \frac{(2M+1)C}{M^2} = \mathcal{O}\left(\frac{1}{M}\right). \end{aligned}$$

Now we consider the estimate

$$\begin{aligned} \left| R_n^{(M)}(x_j) - R_n(x_j) \right| &\leq \sum_{|m|>M} |\phi_{n-m}\widetilde{g}_m| \leq \widetilde{C} \sum_{|m|>M} \left| \frac{1}{n-m}\frac{1}{m^2} \right| \\ &= \widetilde{C}\left(\sum_{m>M} \frac{1}{(m-n)m^2} + \sum_{m>M} \frac{1}{(m+n)m^2} \right) \\ &\overset{|n|\leq M}{\leq} 2\widetilde{C} \sum_{m>M} \frac{1}{(m-M)m^2} \\ &\overset{(5.1)}{\leq} 2\widetilde{C}\left(\frac{1}{(M+1)^2} + \int_{M+1}^{\infty} \frac{1}{(x-M)x^2}\, dx \right). \end{aligned}$$

With the primitive integral

$$\int \frac{1}{(x-M)x^2}\, dx = \frac{\ln(x-M)}{M^2} - \frac{\ln(x)}{M^2} + \frac{1}{Mx}$$

we obtain

$$\left| R_n^{(M)}(x_j) - R_n(x_j) \right| = \mathcal{O}\left(\frac{\ln(M)}{M^2}\right). \tag{5.4}$$

This leads to

$$\begin{aligned} \left| R^{(M)}(x; x_j) - R_M(x; x_j) \right| &\leq \sum_{|n|\leq M} \left| R_n^{(M)}(x_j) - R_n(x_j) \right|, \\ &\leq \sum_{|n|\leq M} \frac{C\ln(M)}{M^2} \\ &= \frac{(2M+1)C\ln(M)}{M^2} \end{aligned}$$

and thus

$$\left| R^{(M)}(x; x_j) - R_M(x; x_j) \right| = \mathcal{O}\left(\frac{\ln(M)}{M}\right).$$

Next we consider

$$\begin{aligned}\left|S_n^{(M)}(x_k) - S_n(x_k)\right| &\leq \sum_{|m|>M} \left|\widetilde{f}_{n-m}\phi_m\right| \leq C \sum_{|m|>M} \left|\frac{1}{(n-m)^2}\frac{1}{m}\right| \\ &\leq \frac{C}{M}\sum_{|m|>M}\frac{1}{(n-m)^2} \\ &= \frac{C}{M}\left(\sum_{m>M}\frac{1}{(n-m)^2} + \sum_{m>M}\frac{1}{(n+m)^2}\right).\end{aligned}$$

With Theorem 5.3.2 we obtain

$$\begin{aligned}\sum_{m>M}\frac{1}{(n-m)^2} &\leq \frac{1}{(M+1-|n|)^2} + \sum_{m=M+2}^{\infty}\frac{1}{(n-m)^2} \\ &\leq \frac{1}{(M+1-|n|)^2} + \int_{m=M+1}^{\infty}\frac{1}{(n-x)^2}dx \\ &\leq \frac{1}{(M+1-|n|)^2} + \frac{1}{M+1-|n|} \\ &\leq \frac{2}{M+1-|n|}\end{aligned}$$

and

$$\begin{aligned}\sum_{m>M}\frac{1}{(n+m)^2} &\leq \frac{1}{(M+1-|n|)^2} + \sum_{m=M+2}^{\infty}\frac{1}{(n+m)^2} \\ &\leq \frac{1}{(M+1-|n|)^2} + \int_{m=M+1}^{\infty}\frac{1}{(n+x)^2}dx \\ &\leq \frac{1}{(M+1-|n|)^2} + \frac{1}{M+1-|n|} \\ &\leq \frac{2}{M+1-|n|}.\end{aligned}$$

Altogether this yields

$$\left|S_n^{(M)}(x_k) - S_n(x_k)\right| \leq \frac{4C}{M(M+1-|n|)},$$

which leads to

$$\left|S^{(M)}(x;x_k) - S_M(x;x_k)\right| \leq \sum_{|n|\leq M}\left|S_n^{(M)}(x_k) - S_n(x_k)\right|,$$

$$
\begin{aligned}
&\leq \sum_{n=-M}^{M} \frac{4C}{M(M+1-|n|)} \\
&= \frac{4C}{M}\left(\frac{1}{M+1} + 2\sum_{n=1}^{M}\frac{1}{M+1-n}\right) \\
&= \frac{4C}{M}\left(\frac{1}{M+1} + 2\sum_{k=1}^{M}\frac{1}{k}\right) \\
&= \mathcal{O}\left(\frac{\ln(M)}{M}\right).
\end{aligned}
$$

Finally, we analyze whether the following expression tends to zero as $M \to \infty$:

$$
\begin{aligned}
&T^{(M)}(x; x_j, x_k) - T_M(x; x_j, x_k) \\
&= \sum_{|n|\leq M} \left(T_n^{(M)}(x_j, x_k) - T_n(x_j, x_k)\right) e^{\mathrm{i}nx} \\
&= -\sum_{|n|\leq M}\sum_{|m|>M} \Phi_{n-m}\Phi_m e^{-\mathrm{i}(n-m)x_j} e^{-\mathrm{i}mx_k} e^{\mathrm{i}nx} \\
&= -\sum_{|n|\leq M}\sum_{|m|>M} \frac{e^{-\mathrm{i}(n-m)x_j}}{2\mathrm{i}(n-m)} \frac{e^{-\mathrm{i}mx_k}}{2\mathrm{i}m} e^{\mathrm{i}nx} \\
&= -\frac{1}{4}\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2} + \frac{1}{4}\sum_{0<|n|\leq M}\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}e^{\mathrm{i}n(x-x_j)}}{(n-m)m} \\
&= -\frac{1}{4}\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2} \\
&\quad + \frac{1}{4}\sum_{0<|n|\leq M}\sum_{|m|>M} \left[\frac{e^{\mathrm{i}m(x_j-x_k)}e^{\mathrm{i}n(x-x_j)}}{nm} + \frac{e^{\mathrm{i}m(x_j-x_k)}e^{\mathrm{i}n(x-x_j)}}{n(n-m)}\right] \\
&= -\frac{1}{4}\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2} + \frac{1}{4}\left[\sum_{0<|n|\leq M} \frac{e^{\mathrm{i}n(x-x_j)}}{n}\right]\left[\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m}\right] \\
&\quad + \frac{1}{4}\sum_{0<|n|\leq M} \frac{e^{\mathrm{i}n(x-x_j)}}{n} \sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{n-m}.
\end{aligned}
$$

We will show that, as $M \to \infty$, the expression

$$-\frac{1}{4}\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2}+\frac{1}{4}\left[\sum_{0<|n|\le M}\frac{e^{\mathrm{i}n(x-x_j)}}{n}\right]\left[\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m}\right] \tag{5.5}$$

uniformly tends to zero with respect to x. This means we have to show that for a given $\epsilon > 0$ there exists an integer N_0, independent of x, such that

$$\left|\frac{1}{4}\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2}-\frac{1}{4}\left[\sum_{0<|n|\le M}\frac{e^{\mathrm{i}n(x-x_j)}}{n}\right]\left[\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m}\right]\right|<\epsilon$$

whenever $M \ge N_0$. The expression

$$\left|\frac{1}{4}\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2}\right|$$

tends to zero as $M \to \infty$ since

$$\left|\frac{1}{4}\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2}\right| \le \frac{1}{4}\sum_{|m|>M}\frac{1}{m^2}=\frac{1}{2}\sum_{m>M}\frac{1}{m^2}.$$

Therefore, for any given $\epsilon > 0$, an integer N_1 exists such that

$$\left|\frac{1}{4}\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2}\right|<\epsilon$$

whenever $M > N_1$. Now we show that

$$\frac{1}{4}\left[\sum_{0<|n|\le M}\frac{e^{\mathrm{i}n(x-x_j)}}{n}\right]\left[\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m}\right]$$

uniformly tends to zero with respect to x as $M \to \infty$. We first consider

$$\sum_{0<|n|\le M}\frac{e^{\mathrm{i}n(x-x_j)}}{n}.$$

If $x = x_j$, then

$$\sum_{0<|n|\le M}\frac{e^{\mathrm{i}n(x-x_j)}}{n}=\sum_{0<|n|\le M}\frac{1}{n}=0.$$

Now let $x \neq x_j$. Then it holds

$$\begin{aligned}\sum_{0<|n|\leq M} \frac{e^{\mathrm{i}n(x-x_j)}}{n} &= \sum_{n=1}^{M} \frac{e^{\mathrm{i}n(x-x_j)}}{n} - \sum_{n=1}^{M} \frac{e^{-\mathrm{i}n(x-x_j)}}{n} \\ &= 2\mathrm{i} \sum_{n=1}^{M} \frac{\sin(n(x-x_j))}{n}.\end{aligned}$$

It is a well-known fact (and easy to verify) that the series

$$\sum_{n=1}^{\infty} \frac{\sin(nx)}{n}$$

is the Fourier series of

$$x \mapsto \frac{\pi - x}{2} \quad (0 \leq x < 2\pi).$$

With Theorem 5.3.4 we obtain that

$$\sum_{0<|n|\leq M} \frac{e^{\mathrm{i}n(x-x_j)}}{n}$$

is uniformly bounded for any $x \in \mathbb{R}$ as $M \to \infty$. In order to show (5.5) we have to show that

$$\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m}$$

tends to zero for any combination of the jump locations x_j of f and x_k of g, respectively. But now, similarly as before, for $x_j = x_k$ we get

$$\sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m} = \sum_{|m|>M} \frac{1}{m} = 0,$$

and for $x_j \neq x_k$ we obtain

$$\left| \sum_{|m|>M} \frac{e^{\mathrm{i}m(x_j-x_k)}}{m} \right| = \left| 2\mathrm{i} \sum_{m=M+1}^{\infty} \frac{\sin(n(x_j-x_k))}{m} \right|. \tag{5.6}$$

Due to the convergence of the Fourier series this tends to zero as $M \to \infty$. Since there are only finitely many jump locations, there are finitely many, say ν combinations $x_j - x_k$ of jump locations. Therefore, for a given $\epsilon > 0$, there exist finitely

many integers $N_2, ..., N_{\nu+1}$, for each jump location combination one integer, such that

$$2\left|\sum_{m=M+1}^{\infty}\frac{\sin(n(x_j-x_k))}{m}\right|<\epsilon$$

for $M \geq \max(N_2, ..., N_{\nu+1})$. This means if we choose $N_0 = \max(N_1, N_2, ..., N_{\nu+1})$ we have proved uniform convergence to zero, with respect to x, of

$$-\frac{1}{4}\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m^2}+\frac{1}{4}\left[\sum_{0<|n|\leq M}\frac{e^{\mathrm{i}n(x-x_j)}}{n}\right]\left[\sum_{|m|>M}\frac{e^{\mathrm{i}m(x_j-x_k)}}{m}\right].$$

If we define

$$\Phi_M(z_1,z_2)=\frac{1}{2}\sum_{0<|n|\leq M}\frac{e^{\mathrm{i}nz_1}}{n}\sum_{|m|>M}\frac{e^{\mathrm{i}mz_2}}{m-n}$$

we can write

$$T^{(M)}(x;x_j,x_k)-T_M(x;x_j,x_k)=-\frac{1}{2}\Phi_M(x-x_j,x_j-x_k)+o(1).$$

The case $z_2 \neq 0$ now corresponds to Theorem 5.4.1, because $x_j \neq x_k$ means that f and g do not have concurrent jump locations. We show that for $z_2 \notin \{0, 2\pi\}$ the expression $\Phi_M(z_1, z_2)$ tends to zero as $M \to \infty$. First consider the estimate

$$\begin{aligned}|\Phi_M(z_1,z_2)| &\leq \sum_{0<|n|\leq M}\frac{1}{2|n|}\left|\sum_{|m|>M}\frac{e^{\mathrm{i}mz_2}}{m-n}\right| \\ &\leq \sum_{0<|n|\leq M}\frac{1}{2|n|}\left(\left|\sum_{k=M-n+1}^{\infty}\frac{e^{\mathrm{i}kz_2}}{k}\right|+\left|\sum_{k=M+n+1}^{\infty}\frac{e^{-\mathrm{i}kz_2}}{k}\right|\right).\end{aligned}$$

Now consider

$$\sum_{k=L}^{\infty}\frac{e^{\mathrm{i}kx}}{k},$$

and take the estimate of inequality (5.2), from the proof of Theorem 5.3.6, to realize that

$$\left|\sum_{k=L}^{\infty}\frac{e^{\mathrm{i}kz_2}}{k}\right|\leq\frac{C}{L}=\mathcal{O}\left(\frac{1}{L}\right)\quad(0<z_2<2\pi),$$

with $C = C(z_2) = \frac{\sqrt{8}}{\sqrt{1-\cos(z_2)}}$. Now we have

$$|\Phi_M(z_1,z_2)| \leq \sum_{0<|n|\leq M}\frac{1}{2|n|}\left(\frac{C}{M-n+1}+\frac{C}{M+n+1}\right)$$

$$\begin{aligned}
&= \sum_{n=1}^{M} \frac{1}{n} \left(\frac{C}{M-n+1} + \frac{C}{M+n+1} \right) \\
&< \sum_{n=1}^{M} \frac{2C}{n(M-n+1)} \\
&= \frac{2C}{M+1} \sum_{n=1}^{M} \left(\frac{1}{n} + \frac{1}{M-n+1} \right) \\
&= \frac{4C}{M+1} \sum_{n=1}^{M} \frac{1}{n} \\
&= \mathcal{O}\left(\frac{\ln(M)}{M} \right),
\end{aligned}$$

which yields that $\Phi_M(z_1, z_2)$ tends to zero as $M \to \infty$, for $z_2 \notin \{0, 2\pi\}$. Next, we consider the case which corresponds to Theorem 5.4.2, i.e. $z_2 = 0$, which means that $\Phi_M = \Phi_M(\cdot, 0)$ is as stated in (5.3). We define $\chi_n^{(M)}$ as

$$\chi_n^{(M)} := \frac{1}{n} \sum_{|m|>M} \frac{1}{m-n} \quad (1 \le n \le M).$$

Then, for $1 \le n \le M$ we obtain

$$\begin{aligned}
\chi_n^{(M)} &= \frac{1}{n} \sum_{|m|>M} \frac{1}{m-n} \\
&= \frac{1}{n} \left(\frac{1}{M+1-n} + \frac{1}{M+2-n} + \frac{1}{M+3-n} + ... + \frac{1}{M+n} \right) \\
&= \frac{1}{n} \left(\sum_{m=1}^{2n} \frac{1}{M+m-n} \right),
\end{aligned}$$

and for $1 \le n+1 \le M$ we obtain

$$\begin{aligned}
\chi_{n+1}^{(M)} &= \frac{1}{n+1} \sum_{|m|>M} \frac{1}{m-n-1} \\
&= \frac{1}{n+1} \left(\frac{1}{M-n} + \frac{1}{M+1-n} + \frac{1}{M+2-n} + \frac{1}{M+3-n} + ... + \frac{1}{M+n+1} \right) \\
&= \frac{1}{n+1} \left(\sum_{m=0}^{2n+1} \frac{1}{M+m-n} \right)
\end{aligned}$$

$$= \frac{1}{n+1}\frac{1}{M-n} + \frac{1}{n+1}\left(\sum_{m=1}^{2n}\frac{1}{M+m-n}\right) + \frac{1}{n+1}\frac{1}{M+n+1},$$

which yields the relation

$$\chi_{n+1}^{(M)} = \frac{n}{n+1}\chi_n^{(M)} + \frac{1}{n+1}\left(\frac{1}{M-n} + \frac{1}{M+n+1}\right).$$

This is equivalent to

$$(n+1)\chi_{n+1}^{(M)} = n\chi_n^{(M)} + \frac{1}{M-n} + \frac{1}{M+n+1},$$

and thus

$$n\left(\chi_{n+1}^{(M)} - \chi_n^{(M)}\right) = \frac{1}{M-n} + \frac{1}{M+n+1} - \chi_{n+1}^{(M)}.$$

From this identity it follows that

$$\chi_{n+1}^{(M)} - \chi_n^{(M)} > 0$$

is equivalent to

$$\frac{1}{M-n} + \frac{1}{M+n+1} - \chi_{n+1}^{(M)} > 0. \tag{5.7}$$

Next we prove that (5.7) holds for $1 \le n < M$:

$$\begin{aligned}
&\frac{1}{M-n} + \frac{1}{M+n+1} - \chi_{n+1}^{(M)} > 0 \\
\Leftrightarrow\ &\underbrace{\frac{1}{M-n} + \ldots + \frac{1}{M-n}}_{n+1} + \underbrace{\frac{1}{M+n+1} + \ldots + \frac{1}{M+n+1}}_{n+1} - \sum_{m=0}^{2n+1}\frac{1}{M+m-n} > 0 \\
\Leftrightarrow\ &\underbrace{\frac{1}{M-n} + \ldots + \frac{1}{M-n}}_{n} + \underbrace{\frac{1}{M+n+1} + \ldots + \frac{1}{M+n+1}}_{n} - \sum_{m=1}^{2n}\frac{1}{M+m-n} > 0 \\
\Leftrightarrow\ &\left(\sum_{m=1}^{n}\frac{1}{M-n} + \frac{1}{M+n+1}\right) - \left(\sum_{m=1}^{2n}\frac{1}{M+m-n}\right) > 0 \\
\Leftrightarrow\ &\left(\sum_{m=1}^{n}\frac{1}{M-n} + \frac{1}{M+n+1}\right) - \left(\sum_{m=1}^{n}\frac{1}{M-n+m} + \frac{1}{M+n+1-m}\right) > 0 \\
\Leftrightarrow\ &\left(\sum_{m=1}^{n}\frac{2M+1}{M^2+M-n^2-n}\right) - \left(\sum_{m=1}^{n}\frac{2M+1}{M^2+M-n^2-n+2nm+m-m^2}\right) > 0.
\end{aligned}$$

Now with

$$\frac{2M+1}{M^2+M-n^2-n+2nm+m-m^2} \leq \frac{2M+1}{M^2+M-n^2-n+nm+m}$$

and

$$\frac{2M+1}{M^2+M-n^2-n+nm+m} < \frac{2M+1}{M^2+M-n^2-n}$$

equation (5.7) holds. Applying Abel's partial summation to (5.3) yields

$$\Phi_M(x) = \left(\sum_{l=1}^{M}\cos(lx)\right)\chi_M^{(M)} + \sum_{n=1}^{M-1}\left(\sum_{l=1}^{n}\cos(lx)\right)\left(\chi_n^{(M)} - \chi_{n+1}^{(M)}\right).$$

For $x \neq 0$, i.e. $x \in (0, 2\pi)$ the sum of the cosines is uniformly bounded (independent of M). In order to prove that, we will use the following identity and the resulting inequality

$$\sum_{l=1}^{M}\cos(lx) = \Re\left(\sum_{l=1}^{M}e^{\mathrm{i}lx}\right) \Rightarrow \left|\sum_{l=1}^{M}\cos(lx)\right| \leq \left|\sum_{l=1}^{M}e^{\mathrm{i}lx}\right|.$$

By Lemma 5.3.5 we know

$$\left|\sum_{l=1}^{M}\left(e^{\mathrm{i}x}\right)^l\right| \leq \frac{\sqrt{2}}{\sqrt{1-\cos(x)}} =: c(x).$$

With these estimates, the definition of $\chi_n^{(M)}$ and the asymptotic behaviour of the harmonic series we obtain

$$\begin{aligned}
|\Phi_M(x)| &\leq c(x)\left(\chi_M^{(M)} + \sum_{n=1}^{M-1}\left(\chi_{n+1}^{(M)} - \chi_n^{(M)}\right)\right) \\
&= c(x)\left(2\chi_M^{(M)} - \chi_1^{(M)}\right) \\
&= c(x)\left(\frac{2}{M}\left(\sum_{m=1}^{2M}\frac{1}{m}\right) - \frac{1}{M} - \frac{1}{M+1}\right) \leq \mathcal{O}\left(\frac{\ln(M)}{M}\right).
\end{aligned}$$

This means that for a fixed $x \in (0, 2\pi)$ the error term $\Phi_M(x)$ tends to zero as $M \to \infty$. For the case $x = 0$ we set $\Phi_0(0) := 0$ and consider $\Phi_l(0), \Phi_{l-1}(0)$ for $l \geq 1$:

$$\Phi_l(0) = \sum_{n=1}^{l}\frac{1}{n}\sum_{|m|>l}\frac{1}{m-n} = \sum_{n=1}^{l}\chi_n^{(l)},$$

$$\Phi_{l-1}(0) \;=\; \sum_{n=1}^{l-1}\frac{1}{n}\sum_{|m|>l-1}\frac{1}{m-n}=\sum_{n=1}^{l-1}\chi_n^{(l-1)}.$$

If $l\geq 1$, then for the difference of $\Phi_l(0)$ and $\Phi_{l-1}(0)$ the following holds

$$\begin{aligned}
\Phi_l(0)-\Phi_{l-1}(0) &= \chi_l^{(l)}+\sum_{n=1}^{l-1}\left(\chi_n^{(l)}-\chi_n^{(l-1)}\right)\\
&= \frac{1}{l}\sum_{|m|>l}\frac{1}{m-l}+\sum_{n=1}^{l-1}\frac{1}{n}\left(\sum_{|m|>l}\frac{1}{m-n}-\sum_{|m|>l-1}\frac{1}{m-n}\right)\\
&= \frac{1}{l}\sum_{|m|>l}\frac{1}{m-l}+\sum_{n=1}^{l-1}\frac{1}{n}\left(\sum_{m=1}^{2n}\frac{1}{l+m-n}-\sum_{m=0}^{2n-1}\frac{1}{l+m-n}\right)\\
&= \frac{1}{l}\sum_{|m|>l}\frac{1}{m-l}+\sum_{n=1}^{l-1}\frac{1}{n}\left(\frac{1}{l+n}-\frac{1}{l-n}\right)\\
&= \frac{1}{l}\sum_{n=1}^{2l}\frac{1}{n}-\sum_{n=1}^{l-1}\frac{2}{l^2-n^2}\\
&= \frac{1}{l}\sum_{n=1}^{2l}\frac{1}{n}-\frac{1}{l}\sum_{n=1}^{l-1}\left(\frac{1}{l-n}+\frac{1}{l+n}\right)\\
&= \frac{1}{l}\sum_{n=1}^{2l}\frac{1}{n}-\frac{1}{l}\left(\frac{1}{l-1}+\ldots+1\right)-\frac{1}{l}\left(\frac{1}{l+1}+\ldots+\frac{1}{2l-1}\right)\\
&= \frac{1}{l}\sum_{n=1}^{2l}\frac{1}{n}-\frac{1}{l}\sum_{n=1}^{l-1}\frac{1}{n}-\frac{1}{l}\sum_{n=l+1}^{2l-1}\frac{1}{n}\\
&= \frac{1}{l^2}+\frac{1}{2l^2}\\
&= \frac{3}{2l^2}.
\end{aligned}$$

Now with

$$\Phi_l(0)-\Phi_{l-1}(0)=\frac{3}{2l^2}\quad(l\geq 1),$$

we obtain

$$\Phi_M(0)=\sum_{l=1}^{M}\left(\Phi_l(0)-\Phi_{l-1}(0)\right)=\frac{3}{2}\sum_{l=1}^{M}\frac{1}{l^2}.$$

As M tends to infinity we get

$$\lim_{M\to\infty}\Phi_M(0)=\frac{3}{2}\frac{\pi^2}{6}=\frac{\pi^2}{4}.$$

This completes the proof of Theorem 5.4.1 and Theorem 5.4.2. □

Now we state the theorem which deals with the situation when h is the product of two discontinuous functions f and g, such that h is continuous.

Theorem 5.4.3 ([7], Theorem 4.5)**.** *Let $f \in \boldsymbol{P}$ be such that $f(x) \neq 0$ for all $x \in [0, 2\pi)$. Moreover let $g \in \boldsymbol{P}$ be such that the discontinuities of f and g are **complementary**, i.e. $h = fg$ is continuous. If f satisfies either one of the two following conditions*

a) $\Re(\frac{1}{f})$ does not change sign in $[0, 2\pi)$ and $\inf\limits_{x\in[0,2\pi)} \left|\Re(\frac{1}{f(x)})\right| > 0$,

b) $\Im(\frac{1}{f})$ does not change sign in $[0, 2\pi)$ and $\inf\limits_{x\in[0,2\pi)} \left|\Im(\frac{1}{f(x)})\right| > 0$,

then

$$\lim_{M\to\infty} h^{(M)}(x) = h(x)$$

is valid, provided that the coefficients are given by

$$\textit{(Inverse Rule)} \qquad \widetilde{h}_n^{(M)} = \sum_{m=-M}^{+M} \left(\left[\!\left[\frac{1}{f} \right]\!\right]^{-1} \right)_{nm} g_m \tag{5.8}$$

instead of

$$\textit{(Laurent Rule)} \qquad h_n^{(M)} = \sum_{m=-M}^{+M} f_{n-m} g_m. \tag{5.9}$$

Proof. We first show that for any function f satisfying the conditions stated in the theorem, the following estimate holds:

$$\max_{|n|\leq M} \sum_{m=-M}^{+M} \left| \left(\left[\!\left[\frac{1}{f} \right]\!\right]^{-1} \right)_{nm} \right| \leq \mathcal{O}(\sqrt{M}). \tag{5.10}$$

For the sake of convenience we set

$$A := \left[\!\left[\frac{1}{f} \right]\!\right]. \tag{5.11}$$

Let $\mu_{\min}$ be the smallest eigenvalue of AA^H and $\boldsymbol{u} = (u_{-M}, u_{-M+1}, ..., u_0, ..., u_M)^T$ be a corresponding eigenvector. Suppose that $\boldsymbol{u}$ is normalized such that $\|\boldsymbol{u}\| = \sqrt{\boldsymbol{u}^H \boldsymbol{u}} = 1$. We obtain with the Cauchy-Schwarz inequality

$$\begin{aligned}\left|\boldsymbol{u}^H A^H \boldsymbol{u}\right|^2 &= |\langle A^H\boldsymbol{u},\boldsymbol{u}\rangle|^2 \leq \left\|A^H\boldsymbol{u}\right\|^2 \underbrace{\|\boldsymbol{u}\|^2}_{=1} \\ &= \boldsymbol{u}^H AA^H\boldsymbol{u} = \boldsymbol{u}^H\mu_{\min}\boldsymbol{u} = \mu_{\min}\underbrace{\boldsymbol{u}^H\boldsymbol{u}}_{=1} \\ &= \mu_{\min}.\end{aligned}$$

Since A is a Toeplitz matrix that is generated by the Fourier coefficients of $1/f$ we obtain

$$A = \frac{1}{2\pi}\begin{pmatrix} \int_0^{2\pi}\frac{1}{f(x)}dx & \int_0^{2\pi}\frac{1}{f(x)}e^{\mathrm{i}x}dx & \dots & \int_0^{2\pi}\frac{1}{f(x)}e^{\mathrm{i}2Mx}dx \\ \int_0^{2\pi}\frac{1}{f(x)}e^{-\mathrm{i}x}dx & \int_0^{2\pi}\frac{1}{f(x)}dx & \dots & \int_0^{2\pi}\frac{1}{f(x)}e^{\mathrm{i}(2M-1)x}dx \\ \vdots & \vdots & \ddots & \vdots \\ \int_0^{2\pi}\frac{1}{f(x)}e^{-\mathrm{i}2Mx}dx & \int_0^{2\pi}\frac{1}{f(x)}e^{-\mathrm{i}(2M-1)x}dx & \dots & \int_0^{2\pi}\frac{1}{f(x)}dx \end{pmatrix},$$

$$A^H = \frac{1}{2\pi}\begin{pmatrix} \int_0^{2\pi}\frac{1}{\overline{f}(x)}dx & \int_0^{2\pi}\frac{1}{\overline{f}(x)}e^{\mathrm{i}x}dx & \dots & \int_0^{2\pi}\frac{1}{\overline{f}(x)}e^{\mathrm{i}2Mx}dx \\ \int_0^{2\pi}\frac{1}{\overline{f}(x)}e^{-\mathrm{i}x}dx & \int_0^{2\pi}\frac{1}{\overline{f}(x)}dx & \dots & \int_0^{2\pi}\frac{1}{\overline{f}(x)}e^{\mathrm{i}(2M-1)x}dx \\ \vdots & \vdots & \ddots & \vdots \\ \int_0^{2\pi}\frac{1}{\overline{f}(x)}e^{-\mathrm{i}2Mx}dx & \int_0^{2\pi}\frac{1}{\overline{f}(x)}e^{-\mathrm{i}(2M-1)x}dx & \dots & \int_0^{2\pi}\frac{1}{\overline{f}(x)}dx \end{pmatrix},$$

which leads to

$$\begin{aligned}\boldsymbol{u}^H A^H\boldsymbol{u} &= \bar{u}_{-M}\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{\overline{f}(x)}\left(u_{-M}+u_{-M+1}e^{-\mathrm{i}x}+\dots+u_M e^{-\mathrm{i}2Mx}\right)dx\right) \\ &\quad+\bar{u}_{-M+1}\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{\overline{f}(x)}\left(u_{-M}e^{\mathrm{i}x}+u_{-M+1}+\dots+u_M e^{-\mathrm{i}(2M-1)x}\right)dx\right) \\ &\quad\vdots \\ &\quad+\bar{u}_M\left(\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{\overline{f}(x)}\left(u_{-M}e^{\mathrm{i}2Mx}+u_{-M+1}e^{\mathrm{i}(2M-1)x}+\dots+u_M\right)dx\right).\end{aligned}$$

Remembering that we showed

$$\mu_{\min} \geq \left|\boldsymbol{u}^H A^H\boldsymbol{u}\right|^2,$$

and by defining

$$u_M(x) := \sum_{m=-M}^{+M} u_m e^{\mathrm{i}mx},$$

we obtain

$$\mu_{\min} \geq \frac{1}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 \frac{1}{\overline{f}(x)} dx \right|.$$

If the conditions in a) are satisfied we obtain

$$\begin{aligned} \mu_{\min} &\geq \frac{1}{4\pi^2} \left(\left| \int_0^{2\pi} |u_M(x)|^2 \Re\left(\frac{1}{f(x)}\right) dx \right|^2 + \left| \int_0^{2\pi} |u_M(x)|^2 \Im\left(\frac{1}{f(x)}\right) dx \right|^2 \right) \\ &\geq \frac{1}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 \Re\left(\frac{1}{f(x)}\right) dx \right|^2 \\ &= \frac{1}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 \left| \Re\left(\frac{1}{f(x)}\right) \right| dx \right|^2. \end{aligned}$$

Similarly, if the conditions in b) are satisfied we obtain

$$\begin{aligned} \mu_{\min} &\geq \frac{1}{4\pi^2} \left(\left| \int_0^{2\pi} |u_M(x)|^2 \Re\left(\frac{1}{f(x)}\right) dx \right|^2 + \left| \int_0^{2\pi} |u_M(x)|^2 \Im\left(\frac{1}{f(x)}\right) dx \right|^2 \right) \\ &\geq \frac{1}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 \Im\left(\frac{1}{f(x)}\right) dx \right|^2 \\ &= \frac{1}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 \left| \Im\left(\frac{1}{f(x)}\right) \right| dx \right|^2. \end{aligned}$$

Now define

$$K_1 := \inf_{x \in [0,2\pi)} \left| \Re\left(\frac{1}{f(x)}\right) \right| > 0, \quad K_2 := \inf_{x \in [0,2\pi)} \left| \Im\left(\frac{1}{f(x)}\right) \right| > 0.$$

Then we either obtain

$$\mu_{\min} \geq \frac{K_1}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 dx \right|^2,$$

or

$$\mu_{\min} \geq \frac{K_2}{4\pi^2} \left| \int_0^{2\pi} |u_M(x)|^2 dx \right|^2.$$

Due to Parseval's formula [70] the following identity holds

$$\frac{1}{2\pi} \int_0^{2\pi} |u_M(x)|^2 dx = 1.$$

Therefore, if condition a) or b) stated in the theorem is satisfied, a constant $b > 0$, independent of M, exists such that $\mu_{\min} \geq b > 0$. As a consequence the matrix A is invertible and the maximal eigenvalue of $(AA^H)^{-1} = A^{-H}A^{-1}$ is $1/\mu_{\min} \leq 1/b < \infty$. Thus, the spectral norm of A^{-1} is $\|A^{-1}\|_2 = \sqrt{1/\mu_{\min}}$. Considering equation (5.10) we see that the left hand side is $\|A^{-1}\|_\infty$. Due to the fact that for any matrix $B \in \mathbb{C}^{n\times n}$ (check for example [44]) the relation

$$\|B\|_\infty \leq \sqrt{n}\,\|B\|_2$$

holds, it follows that

$$\left\|A^{-1}\right\|_\infty \leq \sqrt{2M+1}\left\|A^{-1}\right\|_2 \leq \sqrt{\frac{2M+1}{b}} = \mathcal{O}(\sqrt{M}),$$

i.e. we have proved (5.10).

Remembering that $h = fg$ it follows that $g = (1/f)h$. Since we assumed that the discontinuities of f and g are complementary, h is continuous. With the estimates of the previous theorem it follows

$$\sum_{m=-M}^{M}\left(\frac{1}{f}\right)_{n-m} h_m = g_n + \underbrace{\mathcal{O}\left(\frac{\ln(M)}{M^2}\right)}_{:=\delta_n}. \tag{5.12}$$

The term δ_n is meant to be the error term in equation (5.4). If $h_n^{(M)}$ is given by the inverse rule and $x \in [0, 2\pi)$ it follows that

$$\begin{aligned}
h_n - h_n^{(M)} &= h_n - \sum_{m=-M}^{M}\left(\left[\!\left[\frac{1}{f}\right]\!\right]^{-1}\right)_{nm} g_m \\
&\overset{(5.12)}{=} \sum_{m=-M}^{M}\left(\left[\!\left[\frac{1}{f}\right]\!\right]^{-1}\right)_{nm} g_m + \sum_{m=-M}^{M}\left(\left[\!\left[\frac{1}{f}\right]\!\right]^{-1}\right)_{nm}\delta_m \\
&\quad - \sum_{m=-M}^{M}\left(\left[\!\left[\frac{1}{f}\right]\!\right]^{-1}\right)_{nm} g_m \\
&= \sum_{m=-M}^{M}\left(\left[\!\left[\frac{1}{f}\right]\!\right]^{-1}\right)_{nm}\delta_m
\end{aligned}$$

and finally

$$\begin{aligned}\left|h^M(x) - h_M(x)\right| &= \sum_{n=-M}^{M}\sum_{m=-M}^{M}\left|\left(\left[\!\left[\frac{1}{f}\right]\!\right]^{-1}\right)_{nm}\right||\delta_m| \\ &\leq (2M+1)\frac{C\sqrt{M}\ln(M)}{M^2} = \mathcal{O}\left(\frac{\ln(M)}{\sqrt{M}}\right).\end{aligned}$$

□

5.5 Examples

In this section we consider examples which illustrate the Fourier factorization theorems. The examples will also illustrate, that the application of the factorization rules are strict in that sense, that the inverse rule in general cannot be applied in the case when it is allowed to use the Laurent rule.

5.5.1 Continuous with discontinuous

We define the functions $f, g : [0, 2\pi) \to \mathbb{R}$ by

$$f(x) := \begin{cases} 1, & 0 \leq x < \pi, \\ 5, & \pi \leq x < 2\pi, \end{cases}$$

and

$$g(x) := -\frac{x^4}{\pi^4} + \frac{4x^3}{\pi^3} - \frac{4x^2}{\pi^2} + 3.$$

The functions f and g are considered to be periodically extended. Then we have $h = fg$ with the periodically extended function

$$h(x) := \begin{cases} -\dfrac{x^4}{\pi^4} + \dfrac{4x^3}{\pi^3} - \dfrac{4x^2}{\pi^2} + 3, & 0 \leq x < \pi, \\ -\dfrac{5x^4}{\pi^4} + \dfrac{20x^3}{\pi^3} - \dfrac{20x^2}{\pi^2} + 15, & \pi \leq x < 2\pi. \end{cases}$$

The periodic extension of the function f is discontinuous, the one of g is continuous. According to Li's theorems we have to apply the Laurent rule for the computation of the coefficients of h. In the following we see examples computed for the truncation order $M = 50$.

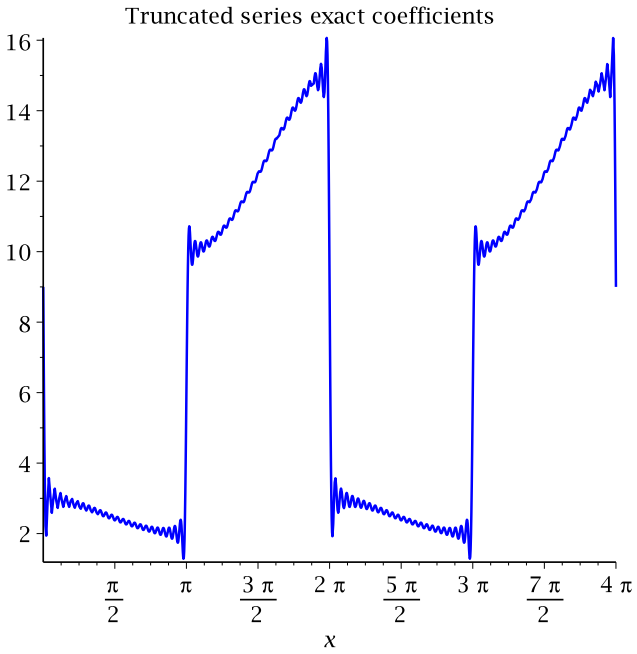


Figure 5.1: Truncated series at $M = 50$ with exact coefficients.

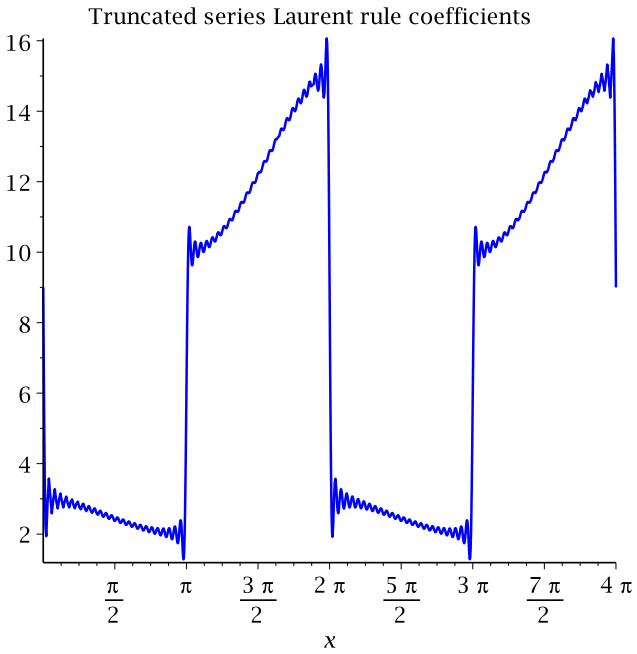


Figure 5.2: Truncated series with coefficients computed with Laurent rule.

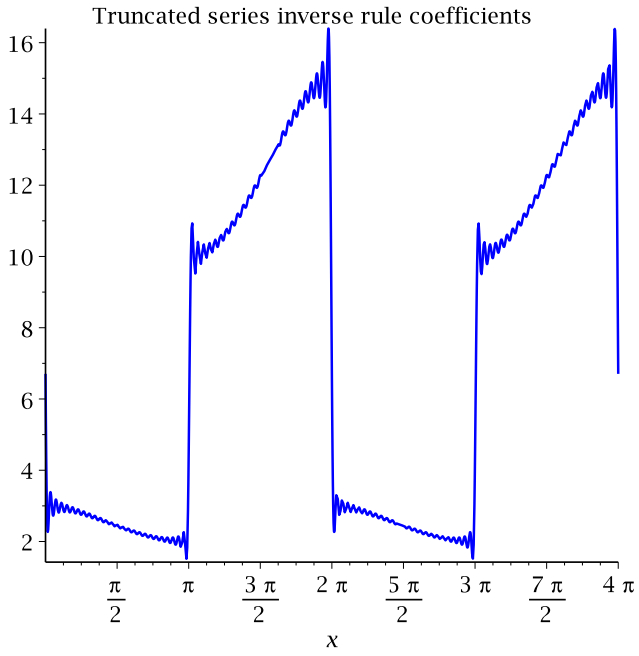


Figure 5.3: Truncated series with coefficients computed with inverse rule.

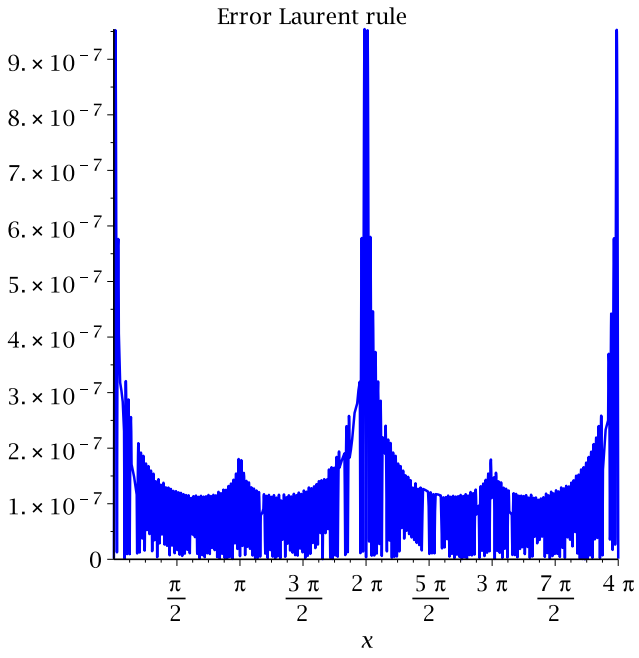


Figure 5.4: Error between truncated series with exact and Laurent rule coefficients.

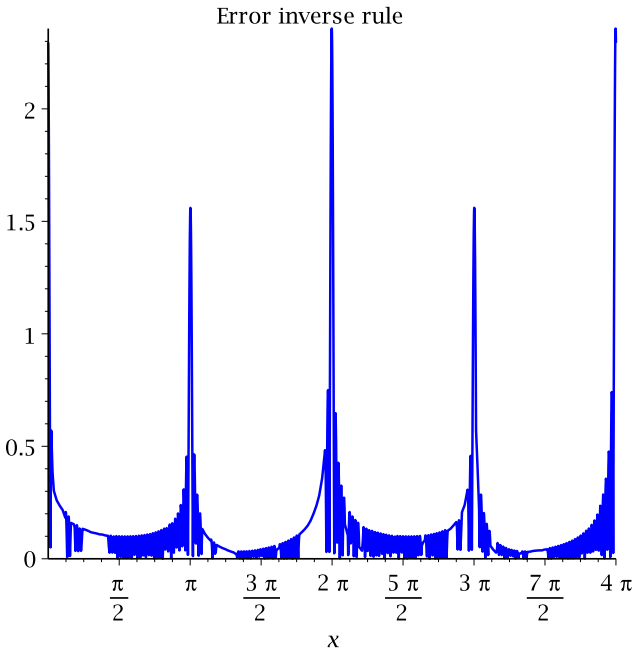


Figure 5.5: Error between truncated series with exact and inverse rule coefficients.

5.5.2 Discontinuous with discontinuous

We define the functions $f, g : [0, 2\pi) \to \mathbb{R}$ by

$$f(x) := \begin{cases} 1, & 0 \le x < \pi, \\ \frac{1}{5}, & \pi \le x < 2\pi, \end{cases}$$

and

$$g(x) := \begin{cases} -\dfrac{x^4}{\pi^4} + \dfrac{4x^3}{\pi^3} - \dfrac{4x^2}{\pi^2} + 3, & 0 \le x < \pi, \\ -\dfrac{5x^4}{\pi^4} + \dfrac{20x^3}{\pi^3} - \dfrac{20x^2}{\pi^2} + 15, & \pi \le x < 2\pi. \end{cases}$$

The functions f and g are considered to be periodically extended. Then we have $h = fg$ with the periodically extended function

$$h(x) := -\frac{x^4}{\pi^4} + \frac{4x^3}{\pi^3} - \frac{4x^2}{\pi^2} + 3.$$

The periodic extensions of the functions f and g are discontinuous, but the one of $h = fg$ is continuous. According to Li's theorems we have to apply the inverse rule for the computation of the coefficients of h.

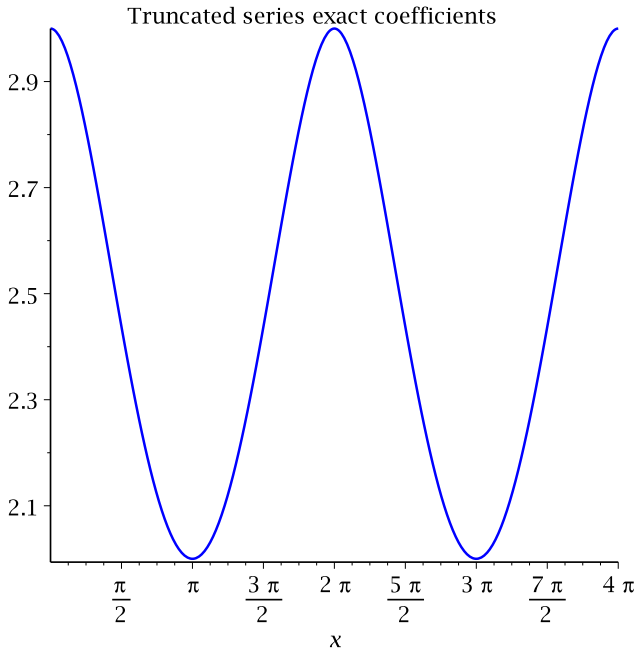


Figure 5.6: Truncated series at $M = 50$ with exact coefficients.

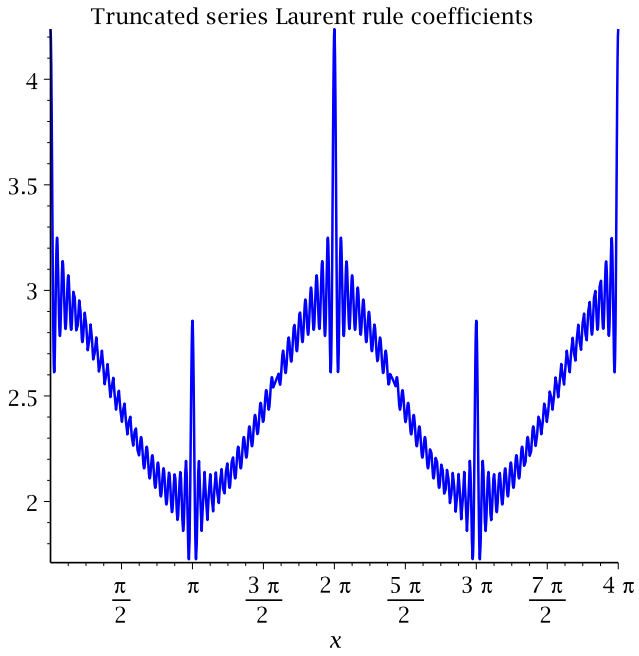


Figure 5.7: Truncated series with coefficients computed with Laurent rule.

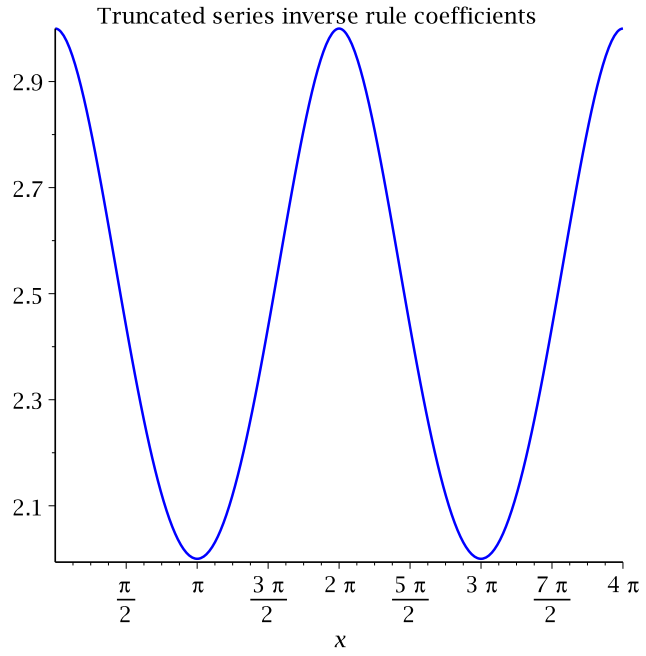


Figure 5.8: Truncated series with coefficients computed with inverse rule.

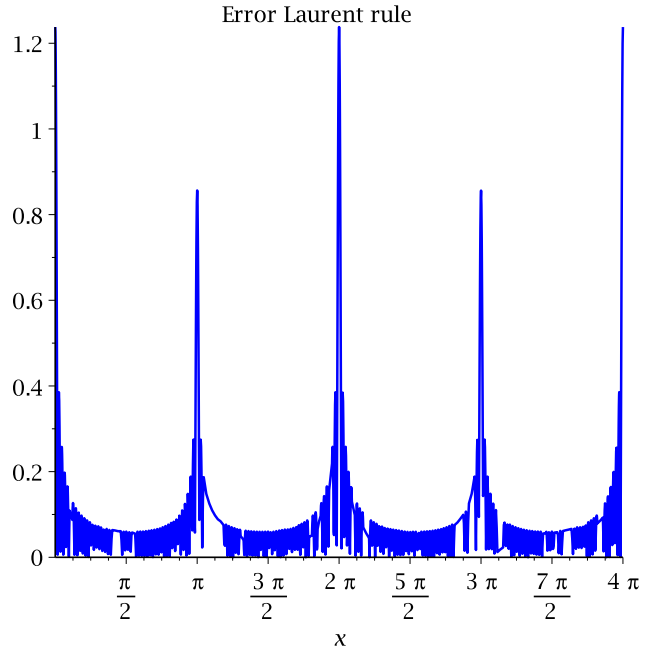


Figure 5.9: Error between truncated series with exact and Laurent rule coefficients.

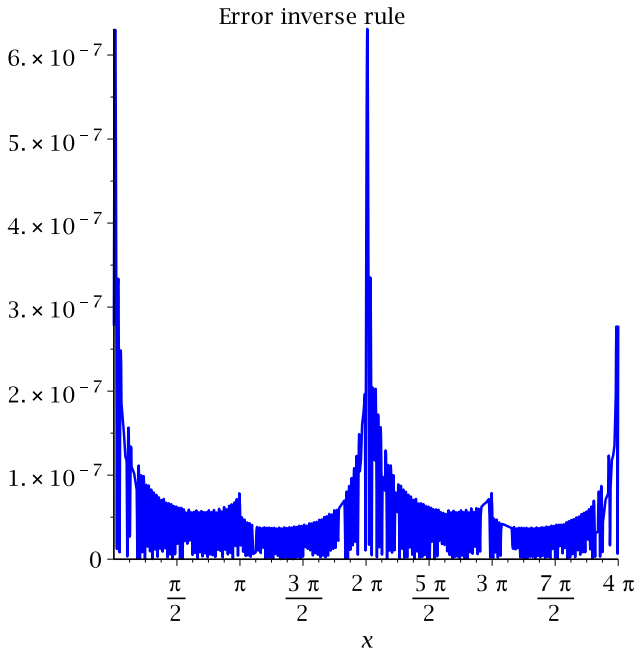


Figure 5.10: Error between truncated series with exact and inverse rule coefficients.

Chapter 6

Helmholtz Problem

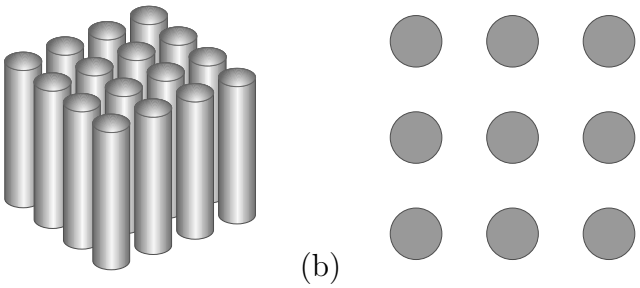
In this chapter we want to discuss how we can solve the *Helmholtz equation* in 2D and 3D for periodic coefficients ε. First we will discuss the 2D case, which will be introduced by simplifying, for 2D photonic crystals, the time-harmonic Maxwell equations to the 2D Helmholtz equation. By an appropriate formulation of the discretized problem the computation of the eigenvalues can be done very efficiently in MATLAB®. With standard built in tools for eigenvalue computations we can compute approximate eigenvalues via a function handle that computes matrix-vector products efficiently, for the matrix whose eigenvalues we are interested in. After that we will see that the same technique can also be used for the 3D Helmholtz equation. Though this case does not correspond to a photonic crystal band structure computation we discuss this case because for periodic coefficients the same techniques can be applied.

6.1 2D Helmholtz equation

6.1.1 2D photonic crystals

In this section we consider structures that are periodic in two spatial directions, however are homogeneous in a third direction. We introduced such structures as 2D photonic crystals. In Figure 6.1 we see an example for 2D photonic crstal consisiting of circular rods. Here we only want to introduce the most important facts. For a much more detailed discussion of the topic we refer to [55].

The function which represents the photonic crystal, namely ε, is a mapping from $\mathbb{R}^2$ to $\mathbb{R}_{>0}$. Therefore we also assume that the electric and the magnetic field of a wave propagating inside the medium also depends only on two spatial variables. Here we consider two different *polarizations* as in [55], namely the so-called *TM-polarization* and *TE-polarization*. A TM-polarized wave is of the form



(a) (b)

Figure 6.1: Example of a two-dimensional photonic crystal with a cylindric rod structure. (a) lateral view (b) view from above

$$\boldsymbol{E}(t,\boldsymbol{x}) = \begin{pmatrix} 0 \\ 0 \\ u(x_1,x_2)e^{\mathrm{i}\omega t} \end{pmatrix}. \tag{6.1}$$

The third identity in (3.4) yields

$$\begin{pmatrix} \partial_2 E_3 \\ -\partial_1 E_3 \\ 0 \end{pmatrix} = -\partial_t \boldsymbol{H},$$

and thus

$$\boldsymbol{H}(t,\boldsymbol{x}) = -\frac{1}{\mathrm{i}\omega}\begin{pmatrix} \partial_2 E_3 \\ -\partial_1 E_3 \\ 0 \end{pmatrix}. \tag{6.2}$$

With equation (6.2) plugged in into the first equation of (3.4), we obtain a Laplace-type scalar eigenvalue problem, the 2D *Helmholtz equation*:

$$-\Delta u(\boldsymbol{x}) = \omega^2 \varepsilon(\boldsymbol{x}) u(\boldsymbol{x}) \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^2. \tag{6.3}$$

Next we consider the case of a TE-polarized wave:

$$\boldsymbol{H}(t,\boldsymbol{x}) = \begin{pmatrix} 0 \\ 0 \\ u(x_1,x_2)e^{\mathrm{i}\omega t} \end{pmatrix}. \tag{6.4}$$

In a similar way as for the TM-polarization case, with the first identity in (3.4) we obtain

$$\boldsymbol{E}(t,\boldsymbol{x}) = \frac{1}{\mathrm{i}\omega\varepsilon}\begin{pmatrix} \partial_2 H_3 \\ -\partial_1 H_3 \\ 0 \end{pmatrix}. \tag{6.5}$$

With this identity and the third identity in (3.4) we obtain a divergence-type scalar eigenvalue problem:

$$-\nabla \cdot \left(\frac{1}{\varepsilon(\boldsymbol{x})} \nabla u(\boldsymbol{x}) \right) = \omega^2 u(\boldsymbol{x}) \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^2. \tag{6.6}$$

6.1.2 Floquet transform

In the previous subsection we have derived scalar eigenvalue problems on whole of $\mathbb{R}^2$. Similar as in section 3.4 for the vectorial 3D problem, we want to reduce the scalar 2D problems to a family of eigenproblems on Ω. For $g \in L^2(\mathbb{R}^2)$ we define, see for example [21], the *Floquet transform*

$$(\mathcal{F}g)(\boldsymbol{k}, \boldsymbol{x}) = e^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \sum_{\boldsymbol{n}\in\mathbb{Z}^2} g(\boldsymbol{x} - \boldsymbol{n}) e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{n}} \qquad \text{for } \boldsymbol{k} \in B. \tag{6.7}$$

As in [37] one can interpret the sum as a Fourier series in the so-called *quasimomentum variable* $\boldsymbol{k}$, with values in $L^2(\Omega)$. Due to the fact that

$$\mathcal{F}(\nabla g) = (\nabla + \mathrm{i}\boldsymbol{k})\mathcal{F}g$$

formally holds, the problem (6.3) is being transformed to

$$-(\nabla + \mathrm{i}\boldsymbol{k}) \cdot (\nabla + \mathrm{i}\boldsymbol{k}) u(\boldsymbol{x}) = \omega^2 \varepsilon(\boldsymbol{x}) u(\boldsymbol{x}) \qquad \text{for } \boldsymbol{x} \in \Omega,\ \boldsymbol{k} \in B. \tag{6.8}$$

Similarly the problem (6.6) is being transformed to

$$-(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \left(\frac{1}{\varepsilon(\boldsymbol{x})} (\nabla + \mathrm{i}\boldsymbol{k}) u(\boldsymbol{x}) \right) = \omega^2 u(\boldsymbol{x}) \qquad \text{for } \boldsymbol{x} \in \Omega,\ \boldsymbol{k} \in B. \tag{6.9}$$

6.1.3 Fourier-Galerkin discretization

We consider the parametrized eigenvalue problem (6.8), subject to periodic boundary conditions, for a fixed $\mathbf{0} \neq \boldsymbol{k} \in B$. Variationally formulated this problem reads: Find $\lambda \in \mathbb{R}$ and $0 \neq u \in H^1_{\mathrm{per}}(\Omega)$ such that

$$a_{\boldsymbol{k}}(u, v) = \lambda b_\varepsilon(u, v) \qquad \text{for all } v \in H^1_{\mathrm{per}}(\Omega), \tag{6.10}$$

where

$$a_{\boldsymbol{k}}(u, v) := \int_\Omega (\nabla + \mathrm{i}\boldsymbol{k}) u(\boldsymbol{x}) \cdot \overline{(\nabla + \mathrm{i}\boldsymbol{k}) v(\boldsymbol{x})}\, d\boldsymbol{x} \tag{6.11}$$

and

$$b_\varepsilon(u, v) := \int_\Omega \varepsilon(\boldsymbol{x}) u(\boldsymbol{x}) \overline{v(\boldsymbol{x})}\, d\boldsymbol{x}, \tag{6.12}$$

with a $\mathbb{Z}^2$-periodic ε. Since we want to do numerical approximation in finite dimesional subspaces with plane waves, for $N \in \mathbb{N}$ we choose the trigonometric function space $\mathcal{T}_N$ as finite dimensional subspace of $H^1_{\text{per}}(\Omega)$. Any element in $\mathcal{T}_N$ can be represented as

$$u_N(\boldsymbol{x}) := \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}. \tag{6.13}$$

In order to transform the weak form of the eigenvalue problem (6.10) into a matrix eigenvalue problem we need to consider the equation

$$a_{\boldsymbol{k}}(u_N, v_N) = \lambda b_\varepsilon(u_N, v_N) \qquad \text{for all } v_N \in \mathcal{T}_N,$$

which we test with all the basis functions of $\mathcal{T}_N$. This means we have to consider

$$a_{\boldsymbol{k}}\big(u_N, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) = \lambda_N b_\varepsilon\big(u_N, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big)$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. Due to linearity we obtain

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} a_{\boldsymbol{k}}\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) = \lambda_N \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} b_\varepsilon\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big)$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. With

$$\begin{aligned} a_{\boldsymbol{k}}\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) &= \int_\Omega (\mathrm{i}2\pi\boldsymbol{n} + \mathrm{i}\boldsymbol{k}) e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \cdot (-\mathrm{i}2\pi\boldsymbol{m} - \mathrm{i}\boldsymbol{k}) e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= (\mathrm{i}2\pi\boldsymbol{n} + \mathrm{i}\boldsymbol{k}) \cdot (-\mathrm{i}2\pi\boldsymbol{m} - \mathrm{i}\boldsymbol{k}) \int_\Omega e^{\mathrm{i}2\pi(\boldsymbol{n}-\boldsymbol{m})\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= |2\pi\boldsymbol{m} + \boldsymbol{k}|^2 \delta_{\boldsymbol{m}\boldsymbol{n}}, \end{aligned}$$

where δ is the usual Kronecker-δ, and

$$\begin{aligned} b_\varepsilon\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) &= \int_\Omega \varepsilon(\boldsymbol{x}) e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= \int_\Omega \varepsilon(\boldsymbol{x}) e^{-\mathrm{i}2\pi(\boldsymbol{m}-\boldsymbol{n})\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= \widehat{\varepsilon}_{\boldsymbol{m}-\boldsymbol{n}} \end{aligned}$$

this becomes

$$|2\pi\boldsymbol{m} + \boldsymbol{k}|^2 \widehat{u}_{\boldsymbol{m}} = \lambda_N \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\varepsilon}_{\boldsymbol{m}-\boldsymbol{n}} \widehat{u}_{\boldsymbol{n}},$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. Now we can formulate the discretized problem as a generalized matrix eigenvalue problem

$$D(\boldsymbol{k})\widehat{\boldsymbol{u}} = \lambda_N \, [\![\varepsilon]\!] \, \widehat{\boldsymbol{u}}, \tag{6.14}$$

where $[\![\varepsilon]\!]$ is a BTTB matrix generated by the Fourier coefficients of ε and $D(\boldsymbol{k})$ is a diagonal matrix depending on the parameter $\boldsymbol{k} \in B \setminus \{\mathbf{0}\}$. Due to the lexicographic ordering of our indices we obtain the following structure for the matrix $[\![\varepsilon]\!]$ (with the four Toeplitz blocks in the corners)

$$[\![\varepsilon]\!] := \begin{pmatrix} \widehat{\varepsilon}_{\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{0}} & & & & \widehat{\varepsilon}_{\binom{0}{-2N}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{-2N}} \\ \vdots & \ddots & \vdots & \cdots & \cdots & \cdots & \vdots & \ddots & \vdots \\ \widehat{\varepsilon}_{\binom{2N}{0}} & \cdots & \widehat{\varepsilon}_{\binom{0}{0}} & & & & \widehat{\varepsilon}_{\binom{2N}{-2N}} & \cdots & \widehat{\varepsilon}_{\binom{0}{-2N}} \\ & \vdots & & \ddots & & & & \vdots & \\ & \vdots & & & \ddots & & & \vdots & \\ & \vdots & & & & \ddots & & \vdots & \\ \widehat{\varepsilon}_{\binom{0}{2N}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{2N}} & & & & \widehat{\varepsilon}_{\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{0}} \\ \vdots & \ddots & \vdots & \cdots & \cdots & \cdots & \vdots & \ddots & \vdots \\ \widehat{\varepsilon}_{\binom{2N}{2N}} & \cdots & \widehat{\varepsilon}_{\binom{0}{2N}} & & & & \widehat{\varepsilon}_{\binom{2N}{0}} & \cdots & \widehat{\varepsilon}_{\binom{0}{0}} \end{pmatrix},$$

and the following structure for the diagonal matrix

$$D(\boldsymbol{k}) := \begin{pmatrix} \left|2\pi\binom{-N}{-N} + \boldsymbol{k}\right|^2 & & & & & & & \\ & \ddots & & & & & & \\ & & \left|2\pi\binom{N}{-N} + \boldsymbol{k}\right|^2 & & & & & \\ & & & \ddots & & & & \\ & & & & \ddots & & & \\ & & & & & \left|2\pi\binom{-N}{N} + \boldsymbol{k}\right|^2 & & \\ & & & & & & \ddots & \\ & & & & & & & \left|2\pi\binom{N}{N} + \boldsymbol{k}\right|^2 \end{pmatrix}.$$

Notice that $D(\boldsymbol{k})$ is invertible for $\boldsymbol{k} \neq \mathbf{0}$. Now instead of considering equation (6.14) and computing the smallest eigenvalues, we consider the equation

$$D^{-1}(\boldsymbol{k}) \, [\![\varepsilon]\!] \, \widehat{\boldsymbol{u}} = \frac{1}{\lambda_N} \widehat{\boldsymbol{u}}, \tag{6.15}$$

and compute the largest eigenvalues. This can be done iteratively with standard eigenvalue solvers implemented in MATLAB®. For this purpose we need an efficient way to compute the matrix-vector product $D^{-1}(\boldsymbol{k}) \, [\![\varepsilon]\!] \, \boldsymbol{x}$ for given $\boldsymbol{x}$. In

Section 4.2 we have discussed concepts for a fast Toeplitz product via FFT. So we can compute $\boldsymbol{y} := [\![\varepsilon]\!]\,\boldsymbol{x}$ and afterwards $D^{-1}(\boldsymbol{k})\boldsymbol{y}$. Since $D^{-1}(\boldsymbol{k})$ is a diagonal matrix the operation $D^{-1}(\boldsymbol{k})\boldsymbol{y}$ can be realized as a pointwise product of two vectors. If $d := (2N+1)^2$ is the dimension of the matrix eigenvalue problem we obtain a total computational cost of $\mathcal{O}(d\log(d))$ operations for such a matrix-vector product.

Next we consider the discretization of the problem (6.9), subject to periodic boundary conditions, for a fixed $\boldsymbol{k} \in B\backslash\{\boldsymbol{0}\}$. Variationally formulated this problem reads: Find $\lambda \in \mathbb{R}$ and $0 \neq u \in H^1_{\mathrm{per}}(\Omega)$ such that

$$a_{\boldsymbol{k}}(u,v) = \lambda b(u,v) \qquad \text{for all } v \in H^1_{\mathrm{per}}(\Omega), \tag{6.16}$$

where

$$a_{\boldsymbol{k}}(u,v) := \int_\Omega \frac{1}{\varepsilon(\boldsymbol{x})}(\nabla + \mathrm{i}\boldsymbol{k})u(\boldsymbol{x}) \cdot \overline{(\nabla + \mathrm{i}\boldsymbol{k})v(\boldsymbol{x})}\,d\boldsymbol{x} \tag{6.17}$$

and

$$b(u,v) := \int_\Omega u(\boldsymbol{x})\overline{v(\boldsymbol{x})}\,d\boldsymbol{x}. \tag{6.18}$$

In order to transform the weak form of the eigenvalue problem (6.16) into a matrix eigenvalue problem we need to consider the equation

$$a_{\boldsymbol{k}}(u_N,v_N) = \lambda b(u_N,v_N) \qquad \text{for all } v_N \in \mathcal{T}_N,$$

which we test with all the basis functions of $\mathcal{T}_N$. This means we have to consider

$$a_{\boldsymbol{k}}\big(u_N, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) = \lambda_N b\big(u_N, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big)$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. Due to linearity we obtain

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} a_{\boldsymbol{k}}\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) = \lambda_N \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} b\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big)$$

for all $\boldsymbol{m} \in \mathbb{I}_N$. With

$$\begin{aligned} a_{\boldsymbol{k}}\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) &= \int_\Omega \frac{1}{\varepsilon(\boldsymbol{x})}(\mathrm{i}2\pi\boldsymbol{n} + \mathrm{i}\boldsymbol{k})e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \cdot (-\mathrm{i}2\pi\boldsymbol{m} - \mathrm{i}\boldsymbol{k})e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\,d\boldsymbol{x} \\ &= (\mathrm{i}2\pi\boldsymbol{n} + \mathrm{i}\boldsymbol{k}) \cdot (-\mathrm{i}2\pi\boldsymbol{m} - \mathrm{i}\boldsymbol{k}) \int_\Omega \frac{1}{\varepsilon(\boldsymbol{x})} e^{\mathrm{i}2\pi(\boldsymbol{n}-\boldsymbol{m})\cdot\boldsymbol{x}}\,d\boldsymbol{x} \end{aligned}$$

$$= (2\pi\boldsymbol{n}+\boldsymbol{k})\cdot(2\pi\boldsymbol{m}+\boldsymbol{k})\widehat{\left(\frac{1}{\varepsilon}\right)}_{\boldsymbol{m}-\boldsymbol{n}},$$

and

$$\begin{aligned} b\big(e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\big) &= \int_\Omega e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= \int_\Omega e^{-\mathrm{i}2\pi(\boldsymbol{m}-\boldsymbol{n})\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= \delta_{\boldsymbol{m}\boldsymbol{n}}, \end{aligned}$$

this becomes

$$(2\pi\boldsymbol{m}+\boldsymbol{k})\cdot\sum_{\boldsymbol{n}\in\mathbb{I}_N}(2\pi\boldsymbol{n}+\boldsymbol{k})\widehat{\left(\frac{1}{\varepsilon}\right)}_{\boldsymbol{m}-\boldsymbol{n}}\widehat{u}_{\boldsymbol{n}} = \lambda_N\widehat{u}_{\boldsymbol{m}},$$

for all $\boldsymbol{m}\in\mathbb{I}_N$. The discretized problem can thus be written as a usual matrix eigenvalue problem

$$C\widehat{\boldsymbol{u}} = \lambda\widehat{\boldsymbol{u}}$$

of dimension $d := (2N+1)^2$. However, as we want to use iterative solvers for efficiency reasons we rather decompose, for a given vector $\boldsymbol{x}\in\mathbb{R}^d$, the matrix-vector product $C\boldsymbol{x}$ into three steps. First we define a matrix $A\in\mathbb{R}^{2d\times d}$ which consists of two diagonal blocks stacked below one another, i.e. A is of the form

$$A = \begin{pmatrix} \ddots & \\ \hline & \ddots \end{pmatrix}.$$

More precisely the first diagonal matrix contains the first components of $2\pi\boldsymbol{n}+\boldsymbol{k}$, for $n\in\mathbb{I}_N$, and the second one the second components, respectively. Next we define a block diagonal matrix $E\in\mathbb{C}^{2d\times 2d}$, containing the BTTB matrices generated by the Fourier coefficient of $1/\varepsilon$:

$$E = \begin{pmatrix} [\![\frac{1}{\varepsilon}]\!] & \\ & [\![\frac{1}{\varepsilon}]\!] \end{pmatrix}.$$

Now we can write the matrix-vector product $C\boldsymbol{x}$ as $A^HEA\boldsymbol{x}$. This has the advantage that the matrix A is sparse, so storage and multiplication for A is $\mathcal{O}(d)$. The matrix E consists of two BTTB matrices, so the cost for storage is $\mathcal{O}(d)$ and the computational cost for a matrix-vector product is $\mathcal{O}(d\log(d))$. Unlike for the

Laplace-type problem, now we have a formulation where we need to compute the smallest eigenvalues of the matrix C. So if we want to work with routines built in into MATLAB®, then we need to hand over a function handle that solves for a given right hand side $\boldsymbol{x}$ linear systems with the coefficient matrix C. As mentioned in [21, 35] a diagonal preconditioner can be chosen for such problems. In our numerical examples we will choose the diagonal matrix, whose entries are the diagonal entries of C, as our preconditioner. This means that the preconditioner is a diagonal matrix $D \in \mathbb{C}^{d\times d}$ with the diagonal entries

$$D_{\boldsymbol{nn}} = \widehat{\left(\frac{1}{\varepsilon}\right)}_{\binom{0}{0}} |2\pi\boldsymbol{n} + \boldsymbol{k}|^2.$$

So for the Laplace-type and the divergence-type problem we were able to represent the resulting matrices as a product of sparse and BTTB matrices. We will need those representations in Section 6.2, where we will consider several numerical examples. Before demonstrating the numerical performance with examples, in the next subsection we will first analyze the convergence of the Fourier-Galerkin method for the Helmholtz problem.

6.1.4 Convergence analysis

In [50, 52] the convergence of the Fourier-Galerkin method for a 2D Schrödinger operator with periodic coefficients was analyzed with standard tools for Galerkin methods, which can be found in the standard reference [4] by Babuška and Osborn. Now we want to analyze the convergence of the Fourier-Galerkin discretization of the 2D Helmholtz equation (6.3). We have seen that this equation leads to the $\boldsymbol{k}$-shifted equation (6.8), which is better suited for the discretization because it is formulated on a periodic domain Ω. Therefore, we consider equation (6.8) for the rest of this section. We want to analyze the convergence of the Fourier-Galerkin method for this equation with the same tools as it was presented in [50, 52]. In order to apply the same machinery as in [50, 52] we will need another, yet similar, theorem on the regularity of the solutions. After having proved the appropriate regularity result we can use the same theorems as in [50, 52], in order to obtain the convergence result for the 2D Helmholtz equation. We will obtain the same order of convergence as in [50, 52], which is not surprising since the problems are similar.

Now we consider the weak form of the shifted Helmholtz problem (6.10). This means we want to find $\lambda \in \mathbb{R}$ and $0 \neq u \in H^1_{\text{per}}(\Omega)$ such that

$$a_{\boldsymbol{k}}(u, v) = \lambda b_\varepsilon(u, v) \quad \text{for all } v \in H^1_{\text{per}}(\Omega), \tag{6.19}$$

with $a_{\boldsymbol{k}}(\cdot,\cdot)$ and $b_\varepsilon(\cdot,\cdot)$ defined as in (6.11) and (6.12). The discrete eigenvalue problem is to find $\lambda_N \in \mathbb{R}$ and $0 \neq u_N \in \mathcal{T}_N$ such that

$$a_{\boldsymbol{k}}(u_N, v_N) = \lambda_N b_\varepsilon(u_N, v_N) \quad \text{for all } v_N \in \mathcal{T}_N. \tag{6.20}$$

The weak form of the boundary value problem that corresponds to the underlying differential operator in (6.10), is to find u such that

$$a_{\boldsymbol{k}}(u, v) = b(f, v) \quad \text{for all } v \in H^1_{\text{per}}(\Omega), \tag{6.21}$$

with $a_{\boldsymbol{k}}(\cdot,\cdot)$ as in (6.11) and $b(\cdot,\cdot)$ is defined by

$$b(f, v) = \int_\Omega f\overline{v} d\boldsymbol{x}. \tag{6.22}$$

It is well known, see e.g. [17, 28, 55], that the sesquilinear form $a_{\boldsymbol{k}}$ is Hermitian, bounded and positive semidefinite on $H^1_{\text{per}}(\Omega)$ for all $\boldsymbol{k} \in B$. Moreover, it is known that $a_{\boldsymbol{k}}$ is coercive on $H^1_{\text{per}}(\Omega)$ for all $\boldsymbol{k} \in B\backslash\{\mathbf{0}\}$. It can be shown that for every

$$\varepsilon : \mathbb{R}^2 \to \{\varepsilon_1, \varepsilon_2\},$$

with

$$\varepsilon_1, \varepsilon_2 \in \mathbb{R}, \quad 0 < \varepsilon_1 < \varepsilon_2,$$

and all $\boldsymbol{k} \in B$ an increasing, non-negative sequence of eigenvalues

$$0 \leq \lambda_1 \leq \lambda_2 \leq ... \nearrow \infty$$

to the weak problem (6.19) exists. For $\boldsymbol{k} \neq \mathbf{0}$ the sequence is strictly positive. The corresponding eigenfunctions form a complete system of $H^1_{\text{per}}(\Omega)$, which are orthogonal with respect to b_ε on $L^2_{\text{per}}(\Omega)$. The eigenspaces, that correspond to the eigenvalues, are finite-dimensional and for the j-th smallest eigenvalues we obtain with the Min-Max Principle:

$$\lambda_j = \min_{\substack{U \subset H^1_{\text{per}}(\Omega) \\ \dim(U)=j}} \max_{u \in U\backslash\{0\}} \frac{a_{\boldsymbol{k}}(u,u)}{b_\varepsilon(u,u)} \qquad \text{for all } j \in \mathbb{N}.$$

We have only summarized the main results here, known e.g. from [17, 28, 55]. A detailed discussion of these well-known facts can be found in Chapter 4 of [55]. Not only the two-dimensional case is being discussed in [55], but also the three-dimensional problems that we have introduced in Chapter 3. Now let $\boldsymbol{k} \neq \mathbf{0}$. We define the *solution operator* $T : L^2_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ for the boundary value problem (6.21) by

$$a_{\boldsymbol{k}}(Tf, v) = b(f, v) \quad \text{for all } v \in H^1_{\text{per}}(\Omega). \tag{6.23}$$

Since we consider right hand sides f of the form $f = \varepsilon u$ we moreover define the solution operator $T^{(\varepsilon)} : L^2_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ by

$$a_{\boldsymbol{k}}(T^{(\varepsilon)}u, v) = b_\varepsilon(u, v) \quad \text{for all } v \in H^1_{\text{per}}(\Omega). \tag{6.24}$$

Notice that $(1/\lambda, u)$ is an eigenpair of $T^{(\varepsilon)}$ if and only if (λ, u) is a solution of (6.19). Now we can use the same arguments as in [50, 52]. Since $a_{\boldsymbol{k}}(\cdot,\cdot)$ is bounded, Hermitian and coercive, and $b(\cdot,\cdot)$ is bounded, one can deduce with Lax-Milgram that $T^{(\varepsilon)} : L^2_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ is well defined and bounded. Moreover, it can easily be shown that $T^{(\varepsilon)}$ is self-adjoint with respect to $a_{\boldsymbol{k}}(\cdot,\cdot)$. The mapping $T^{(\varepsilon)} : H^1_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ is also compact since $H^1_{\text{per}}(\Omega) \subset\subset L^2_{\text{per}}(\Omega)$. Moreover, the solution operator $T^{(\varepsilon)}$ is positive definite. From standard spectral theory results it is known that the eigenvalues of the bounded, compact, self-adjoint and positive definite operator $T^{(\varepsilon)}$ are real and positive and can be ordered as

$$0 \swarrow \ldots \leq \frac{1}{\lambda_2} \leq \frac{1}{\lambda_1},$$

counted up to finite multiplicity. The corresponding eigenfunctions

$$u_1,\ u_2, \ldots$$

are orthogonal with respect to $a_{\boldsymbol{k}}(\cdot,\cdot)$ and complete in $L^2_{\text{per}}(\Omega)$. Now we will discuss the regularity results that are needed for the proof of the convergence result. From the Lax-Milgram theorem we know that if u is a solution to (6.21) then $u \in H^1_{\text{per}}(\Omega)$. Another useful theorem is the following. It states that we actually can expect a higher regularity than H^1_{per}-regularity from a solution to a Laplace-type problem. The proof is similar to the non-periodic case which can be found in [54] on page 319. However, our case is even easier to treat because of the periodic boundary conditions.

Theorem 6.1.1. *If $u \in H^1_{\text{per}}(\Omega)$ is a weak solution of the problem (6.21), then $u \in H^2_{\text{per}}(\Omega)$.*

Proof. We will show that for $-\Delta u = f$, with $f \in L^2_{\text{per}}(\Omega)$ and u subject to periodic boundary conditions the identity

$$\int_\Omega |\Delta u|^2 d\boldsymbol{x} = \sum_{i,j=1}^{n} \int_\Omega |u_{x_i x_j}|^2 d\boldsymbol{x}$$

holds, and thus $u \in H^2_{\text{per}}(\Omega)$. Using partial integration and the periodicity of u, and assuming for the moment that u is arbitrarily smooth, we obtain the following:

$$\int_\Omega |\Delta u|^2 d\boldsymbol{x} \;=\; \sum_{i,j=1}^{n} \int_\Omega u_{x_i x_i} u_{x_j x_j} d\boldsymbol{x}$$

$$
\begin{aligned}
&= -\sum_{i,j=1}^{n} \int_{\Omega} u_{x_i} u_{x_j x_i x_j} d\boldsymbol{x} \\
&= \sum_{i,j=1}^{n} \int_{\Omega} u_{x_i x_j} u_{x_i x_j} d\boldsymbol{x} \\
&= \sum_{i,j=1}^{n} \int_{\Omega} |u_{x_i x_j}|^2 d\boldsymbol{x}.
\end{aligned}
$$

With this result we obtain that if u is a solution of (6.21) then $u \in H^2_{\text{per}}(\Omega)$. □

In [50] it was shown that the following theorem (in similar form) holds. It states that the solution u to the boundary value problem (6.21) has a regularity which is two orders higher than that of the right hand side f.

Theorem 6.1.2 ([50], Theorem 3.76)**.** *Let $s \in \mathbb{R}$, with $s \geq 2$, and let $f \in H^{s-2}_{\text{per}}(\Omega)$. Then there exists a unique solution u to the weak problem (6.21) such that $u \in H^s_{\text{per}}(\Omega)$ and*

$$\|u\|_{H^s_{\text{per}}} \leq C \|f\|_{H^{s-2}_{\text{per}}}$$

for a constant $C > 0$.

The next theorem, which can be found in [52], tells us what kind of regularity we can expect for a function, which is a product of two functions in a known regularity class.

Theorem 6.1.3 ([52], Theorem 2.1)**.** *If $s, t \in \mathbb{R}$, $|s| \leq 1$, $t > 1$, $u \in H^t_{\text{per}}(\Omega)$ and $\varepsilon \in H^s_{\text{per}}(\Omega)$, then*

$$\|\varepsilon u\|_{H^s_{\text{per}}} \leq C(t) \|\varepsilon\|_{H^s_{\text{per}}} \|u\|_{H^t_{\text{per}}},$$

where $C(t)$ is a constant independent of ε and u.

Unfortunately, unlike for the Schrödinger equation treated in [52], we will rather need an estimate for products where $u \in H^1_{\text{per}}(\Omega)$. Therefore, next we present a theorem for products of functions in fractional Sobolev spaces which we will need. Due to the equivalence of periodic and usual Sobolev norms, see for example Theorem 3.29 in [50], this result will be also applicable in the periodic setting.

Theorem 6.1.4. *Let $s \in [0, 1/2)$, $\varepsilon \in H^s(\Omega) \cap L^\infty(\Omega)$ and $u \in H^1(\Omega)$. Then there exists a $C > 0$ such that*

$$\|\varepsilon u\|_{H^s} \leq C \|u\|_{H^1}$$

holds, where $C = C(s)$ is independent of u.

Proof. With the definition of the Sobolev-Slobodeckij norm we obtain

$$\|\varepsilon u\|_{H^s}^2 = \|\varepsilon u\|_{W^{s,2}}^2 = \int_\Omega \int_\Omega \frac{|\varepsilon(\boldsymbol{x})u(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})u(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y}.$$

First we rewrite the numerator of the integrand. It holds

$$\varepsilon(\boldsymbol{x})u(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})u(\boldsymbol{y}) = \varepsilon(\boldsymbol{x})(u(\boldsymbol{x}) - u(\boldsymbol{y})) + (\varepsilon(\boldsymbol{x}) - \varepsilon(\boldsymbol{y}))u(\boldsymbol{y}),$$

and therefore we have

$$|\varepsilon(\boldsymbol{x})u(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})u(\boldsymbol{y})|^2 \leq 2\left(|\varepsilon(\boldsymbol{x})|^2|(u(\boldsymbol{x}) - u(\boldsymbol{y}))|^2 + |(\varepsilon(\boldsymbol{x}) - \varepsilon(\boldsymbol{y}))|^2|u(\boldsymbol{y})|^2\right).$$

With this, we obtain

$$\begin{aligned}\|\varepsilon u\|_{H^s}^2 \quad &\leq \quad 2\int_\Omega \int_\Omega |\varepsilon(\boldsymbol{x})|^2 \frac{|u(\boldsymbol{x}) - u(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y} \\ &\quad +2\int_\Omega \int_\Omega |u(\boldsymbol{x})|^2 \frac{|\varepsilon(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y}.\end{aligned}$$

For the first term we obtain

$$\begin{aligned}\int_\Omega \int_\Omega |\varepsilon(\boldsymbol{x})|^2 \frac{|u(\boldsymbol{x}) - u(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y} \quad &\leq \quad \|\varepsilon\|_\infty^2 \int_\Omega \int_\Omega \frac{|u(\boldsymbol{x}) - u(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y} \\ &= \quad \|\varepsilon\|_\infty^2 \|u\|_{H^s}^2 \\ &\lesssim \quad \|u\|_{H^1}^2.\end{aligned}$$

Next we consider the second term. Since $u \in H^1(\Omega)$ we also know that $u \in L^p(\Omega)$ for all $p \in [1, \infty)$. Moreover, we know that $\varepsilon \in L^\infty(\Omega)$. With this we obtain

$$\int_\Omega \int_\Omega |u(\boldsymbol{x})|^2 \frac{|\varepsilon(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y}$$

$$\overset{\text{Hölder}}{\leq} \quad \int_\Omega \left[\int_\Omega |u(\boldsymbol{x})|^{2p} d\boldsymbol{x}\right]^{\frac{1}{p}} \cdot \left[\int_\Omega \frac{|\varepsilon(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})|^{2q}}{|\boldsymbol{x}-\boldsymbol{y}|^{2sq+2q}} d\boldsymbol{x}\right]^{\frac{1}{q}} d\boldsymbol{y}$$

$$\overset{\frac{1}{p}=\frac{2}{2p}}{=} \quad \|u\|_{L^{2p}}^2 \int_\Omega 1 \cdot \left[\int_\Omega \frac{|\varepsilon(\boldsymbol{x}) - \varepsilon(\boldsymbol{y})|^{2q}}{|\boldsymbol{x}-\boldsymbol{y}|^{2sq+2q}} d\boldsymbol{x}\right]^{\frac{1}{q}} d\boldsymbol{y}$$

$$\begin{aligned}
&\overset{\text{Hölder}}{\leq} \quad \|u\|_{L^{2p}}^2 \left[\int_\Omega 1 d\boldsymbol{y}\right]^{\frac{1}{p}} \left[\int_\Omega \left[\int_\Omega \frac{|\varepsilon(\boldsymbol{x})-\varepsilon(\boldsymbol{y})|^{2q}}{|\boldsymbol{x}-\boldsymbol{y}|^{2sq+2q}} d\boldsymbol{x}\right]^{\frac{q}{q}} d\boldsymbol{y}\right]^{\frac{1}{q}} \\
&= \quad \|u\|_{L^{2p}}^2 |\Omega|^{\frac{1}{p}} \left[\int_\Omega \int_\Omega \frac{|\varepsilon(\boldsymbol{x})-\varepsilon(\boldsymbol{y})|^{2q}}{|\boldsymbol{x}-\boldsymbol{y}|^{2sq+2q}} d\boldsymbol{x} d\boldsymbol{y}\right]^{\frac{1}{q}} \\
&\leq \quad \|u\|_{L^{2p}}^2 |\Omega|^{\frac{1}{p}} \left(2\|\varepsilon\|_\infty\right)^{\frac{2(q-1)}{q}} \left[\int_\Omega \int_\Omega \frac{|\varepsilon(\boldsymbol{x})-\varepsilon(\boldsymbol{y})|^{2}}{|\boldsymbol{x}-\boldsymbol{y}|^{2sq+2q}} d\boldsymbol{x} d\boldsymbol{y}\right]^{\frac{1}{q}}.
\end{aligned}$$

Now consider the identity

$$2sq + 2q = 2\widetilde{s} + 2.$$

This is equivalent to

$$\widetilde{s} = sq + q - 1 = q(s+1) - 1.$$

This means that if $s \in [0, 1/2)$ is fixed, then for p large enough we obtain q close enough to 1 such that $\widetilde{s} < 1/2$. Choosing p large enough we obtain

$$\begin{aligned}
&\int_\Omega \int_\Omega |u(\boldsymbol{x})|^2 \frac{|\varepsilon(\boldsymbol{x})-\varepsilon(\boldsymbol{y})|^2}{|\boldsymbol{x}-\boldsymbol{y}|^{2s+2}} d\boldsymbol{x}\, d\boldsymbol{y} \\
&\leq \quad \|u\|_{L^{2p}}^2 |\Omega|^{\frac{1}{p}} \left(2\|\varepsilon\|_\infty\right)^{\frac{2(q-1)}{q}} \left[\int_\Omega \int_\Omega \frac{|\varepsilon(\boldsymbol{x})-\varepsilon(\boldsymbol{y})|^{2}}{|\boldsymbol{x}-\boldsymbol{y}|^{2\widetilde{s}+2}} d\boldsymbol{x} d\boldsymbol{y}\right]^{\frac{1}{q}} \\
&= \quad \|u\|_{L^{2p}}^2 |\Omega|^{\frac{1}{p}} \left(2\|\varepsilon\|_\infty\right)^{\frac{2(q-1)}{q}} \|\varepsilon\|_{W^{\widetilde{s},2}}^{\frac{2}{q}} \\
&\leq \quad C(p) \|u\|_{H^1}^2 |\Omega|^{\frac{1}{p}} \left(2\|\varepsilon\|_\infty\right)^{\frac{2(q-1)}{q}} \|\varepsilon\|_{W^{\widetilde{s},2}}^{\frac{2}{q}} \\
&\lesssim \quad \|u\|_{H^1}^2 .
\end{aligned}$$

□

Let us clarify the question what p large enough in the proof above means. We have to choose p such that

$$\widetilde{s} = q(s+1) - 1 < \frac{1}{2} \quad \Longleftrightarrow \quad q < \frac{3}{2(s+1)}$$

holds. By representing s as $s = 1/2 - \delta$, for some $\delta > 0$, we obtain that p must be chosen such that

$$1 < q < \frac{3}{2\left(\frac{3}{2} - \delta\right)} = \frac{3}{3 - 2\delta}$$

holds. With the relation

$$\frac{1}{p} + \frac{1}{q} = 1 \quad \Longleftrightarrow \quad q = \frac{p}{p-1}$$

we obtain

$$q = \frac{p}{p-1} < \frac{3}{3-2\delta} \quad \Longleftrightarrow \quad p > \frac{3}{2\delta}.$$

Now we proceed in a similar way as in [52]. For the next theorem we need a definition for a class of periodic functions introduced in [52]. We define

$$\mathcal{X}_{\text{per}}(\Omega) := \{f \in H^{1/2-\rho}_{\text{per}}(\Omega) \text{ for any } \rho > 0\} \cap L^\infty(\mathbb{R}^2).$$

We are interested in this class of functions, because the crystal functions ε that we are interested in lie in this class of functions, as it was shown in [52]. Since periodic Sobolev norms are equivalent to usual Sobolev norms (see Theorem 3.29 in [50]), with Theorem 6.1.4 we obtain the following corollary.

Corollary 6.1.5. *Let $\varepsilon \in \mathcal{X}_{\text{per}}(\Omega)$ and $u \in H^1_{\text{per}}(\Omega)$, then*

$$\|\varepsilon u\|_{H^s_{\text{per}}} \lesssim \|u\|_{H^1_{\text{per}}}$$

for $s < 1/2$.

Our goal is to prove the convergence result in the same way as it was done in [52]. Therefore we need the following regularity result for the solution operator $T^{(\varepsilon)}$ defined by (6.24).

Theorem 6.1.6. *Let $\varepsilon \in \mathcal{X}_{\text{per}}(\Omega)$, $u \in H^1_{\text{per}}(\Omega)$ and $f = \varepsilon u$. Then*

$$\left\|T^{(\varepsilon)}u\right\|_{H^{5/2-\rho}_{\text{per}}} \lesssim \|u\|_{H^1_{\text{per}}} \qquad \text{for any } \rho > 0.$$

Proof. With Theorem 6.1.2 and Corollary 6.1.5 we obtain

$$\left\|T^{(\varepsilon)}u\right\|_{H^{s+2}_{\text{per}}} = \|Tf\|_{H^{s+2}_{\text{per}}} \lesssim \|f\|_{H^s_{\text{per}}} = \|\varepsilon u\|_{H^s_{\text{per}}} \lesssim \|u\|_{H^1_{\text{per}}}.$$

□

The next corollary tells us what kind of regularity we can expect for the eigenfunctions of the 2D Helmholtz problem. This result follows from Theorem 6.1.6.

Corollary 6.1.7. *If u is an eigenfunction of (6.19) with $\varepsilon \in \mathcal{X}_{\text{per}}$, then*

$$\|u\|_{H^{5/2-\rho}_{\text{per}}} \lesssim \|u\|_{H^1_{\text{per}}} \qquad \text{for any } \rho > 0.$$

Now we have prepared all regularity results which we need in order to proceed exactly the same way as in [52]. First we define the discrete solution operator $T_N^{(\varepsilon)} : L^2_{\text{per}}(\Omega) \to \mathcal{T}_N$ to (6.21). For $N \in \mathbb{N}$ and $f \in L^2_{\text{per}}(\Omega)$ we define $T_N^{(\varepsilon)}$ by

$$a_{\boldsymbol{k}}(T_N^{(\varepsilon)} f, v_N) = b_\varepsilon(f, v_N) \quad \text{ for all } v_N \in \mathcal{T}_N. \tag{6.25}$$

Then $T_N^{(\varepsilon)}$ is bounded and self-adjoint with respect to $a_{\boldsymbol{k}}(\cdot,\cdot)$. If the projection Q_N is defined by $a_{\boldsymbol{k}}(Q_N u - u, v) = 0$, for all $u \in H^1_{\text{per}}(\Omega)$ and all $v_N \in \mathcal{T}_N$, then the discrete solution operator $T_N^{(\varepsilon)}$ is the projection of the solution operator $T^{(\varepsilon)}$, i.e. $T_N^{(\varepsilon)} = Q_N T^{(\varepsilon)}$. From the compactness of $T^{(\varepsilon)} : H^1_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ it follows that $T_N^{(\varepsilon)} : H^1_{\text{per}}(\Omega) \to H^1_{\text{per}}(\Omega)$ is compact, and any λ_N is an eigenvalue of (6.20) if and only if λ_N^{-1} is an eigenvalue of $T_N^{(\varepsilon)}$. For $s \in \mathbb{R}$ and $N \in \mathbb{N}$ we define the orthogonal projection from $H^s_{\text{per}}(\Omega)$ onto $\mathcal{T}_N$, which actually is the truncated Fourier series, such that for all $u \in H^s_{\text{per}}(\Omega)$ it holds

$$P_N u(\boldsymbol{x}) := \sum_{\boldsymbol{n} \in \mathbb{I}_N} \widehat{u}_{\boldsymbol{n}} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^2. \tag{6.26}$$

The next result and the proof is adopted from [52]. The two results for $T^{(\varepsilon)}$ that follow are the same as the two corresponding results for the Schrödinger problem in [52] (Corollary 4.4 and Lemma 4.5), and they are proved in the same way.

Theorem 6.1.8 ([52], Lemma 4.3)**.** *For $s, t \in \mathbb{R}$ with $s \le t$, and $N \in \mathbb{N}$, if $u \in H^t_{\text{per}}(\Omega)$, then*

$$\|u - P_N u\|_{H^s_{\text{per}}} \le N^{s-t} \|u\|_{H^t_{\text{per}}}.$$

Proof. For $s \le t \in \mathbb{R}$, $u \in H^t_{\text{per}}$, and $N \in \mathbb{N}$ we have

$$\begin{aligned}
\|u - P_N u\|^2_{H^s_{\text{per}}} &= \sum_{|\boldsymbol{n}|>N} |\boldsymbol{n}|^{2s} |\widehat{u}_{\boldsymbol{n}}|^2 \\
&= \sum_{|\boldsymbol{n}|>N} |\boldsymbol{n}|^{2s} ||\boldsymbol{n}|^{-2t} ||\boldsymbol{n}|^{2t} |\widehat{u}_{\boldsymbol{n}}|^2 \\
&\le N^{2s-2t} \sum_{|\boldsymbol{n}|>N} |\boldsymbol{n}|^{2t} |\widehat{u}_{\boldsymbol{n}}|^2 \\
&\le N^{2s-2t} \|u\|^2_{H^t_{\text{per}}}.
\end{aligned}$$

□

Corollary 6.1.9. *Let $\varepsilon \in \mathcal{X}_{\mathrm{per}}$. For $u \in H^1_{\mathrm{per}}(\Omega)$ and $\rho > 0$,*

$$\inf_{\chi \in \mathcal{T}_N} \left\| T^{(\varepsilon)} u - \chi \right\|_{H^1_{\mathrm{per}}} \lesssim N^{-3/2+2\rho} \left\| u \right\|_{H^1_{\mathrm{per}}}.$$

Moreover, if u is an eigenfunction of (6.19) and $\rho > 0$, then

$$\inf_{\chi \in \mathcal{T}_N} \left\| u - \chi \right\|_{H^1_{\mathrm{per}}} \lesssim N^{-3/2+2\rho} \left\| u \right\|_{H^1_{\mathrm{per}}}.$$

Proof. Let $\rho > 0$ and $\chi := P_N T^{(\varepsilon)} u$. Then it follows from Lemmas 6.1.6 and 6.1.8 that

$$\begin{aligned} \inf_{\chi \in \mathcal{T}_N} \left\| T^{(\varepsilon)} u - \chi \right\|_{H^1_{\mathrm{per}}} &\leq \left\| T^{(\varepsilon)} u - P_N T^{(\varepsilon)} u \right\|_{H^1_{\mathrm{per}}} \\ &\leq N^{-3/2+\rho} \left\| T^{(\varepsilon)} u \right\|_{H^{5/2-\rho}_{\mathrm{per}}} \\ &\lesssim N^{-3/2+\rho} \left\| u \right\|_{H^1_{\mathrm{per}}}. \end{aligned}$$

With Lemmas 6.1.6 and 6.1.9 we obtain

$$\begin{aligned} \inf_{\chi \in \mathcal{T}_N} \left\| u - \chi \right\|_{H^1_{\mathrm{per}}} &\leq \left\| u - P_N u \right\|_{H^1_{\mathrm{per}}} \\ &\leq N^{-3/2+\rho} \left\| u \right\|_{H^{5/2-\rho}_{\mathrm{per}}} \\ &\lesssim N^{-3/2+\rho} \left\| u \right\|_{H^1_{\mathrm{per}}}. \end{aligned}$$

□

Lemma 6.1.10. *Let $\varepsilon \in \mathcal{X}_{\mathrm{per}}$. Then*

$$\left\| T^{(\varepsilon)} - T_N^{(\varepsilon)} \right\|_{H^1_{\mathrm{per}}} \lesssim N^{-3/2+\rho} \qquad \text{for any } \rho > 0.$$

Proof. With Cea's lemma (e.g. in [14], Theorem 2.4.1) and Theorem 6.1.9 it follows

$$\begin{aligned} \left\| T^{(\varepsilon)} - T_N^{(\varepsilon)} \right\|_{H^1_{\mathrm{per}}} &= \sup_{\substack{u \in H^1_{\mathrm{per}} \\ \|u\|_{H^1_{\mathrm{per}}} \neq 0}} \frac{\left\| T^{(\varepsilon)} u - T_N^{(\varepsilon)} u \right\|_{H^1_{\mathrm{per}}}}{\left\| u \right\|_{H^1_{\mathrm{per}}}} \\ &\lesssim \sup_{\substack{u \in H^1_{\mathrm{per}} \\ \|u\|_{H^1_{\mathrm{per}}} \neq 0}} \inf_{\chi \in \mathcal{T}_N} \frac{\left\| T^{(\varepsilon)} u - \chi \right\|_{H^1_{\mathrm{per}}}}{\left\| u \right\|_{H^1_{\mathrm{per}}}} \lesssim N^{-3/2+\rho}. \end{aligned}$$

□

Now we define the *gap between two subspaces* of a Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$:

$$\delta_{\mathcal{H}}(X,Y) := \sup_{x\in X, \|x\|_{\mathcal{H}}=1} \operatorname{dist}(x,Y) = \sup_{y\in Y, \|y\|_{\mathcal{H}}=1} \operatorname{dist}(y,X).$$

Now we are ready to state the main theorem for the discretization of the Helmholtz problem with Fourier-Galerkin. It is actually the same convergence result as for the Schrödinger problem considered in [52] (Theorem 4.6).

Theorem 6.1.11. *Let $\varepsilon \in \mathcal{X}_{\mathrm{per}}$ and let λ be an eigenvalue of (6.19) with multiplicity m and corresponding eigenspace M. The for N sufficiently large and $\rho > 0$ arbitrarily small, there exist m eigenvalues $\lambda_1, ..., \lambda_m$ of (6.20) (counted accordingly to their multiplicity) with $\lambda_j = \lambda_j(N)$ and with corresponding eigenspaces $M(\lambda_1), ..., M(\lambda_m) \subset \mathcal{T}_N$ such that*

$$\delta_{H^1_{\mathrm{per}}}(M, \mathcal{M}_N) \lesssim N^{-3/2+\rho} \quad , \qquad \text{where } \mathcal{M}_N := \bigoplus_{j=1}^{m} M_j(\lambda_j), \quad \text{and}$$

$$|\lambda - \lambda_j| \lesssim N^{-3+2\rho} \quad , \qquad \text{for } j = 1, ..., m.$$

This theorem follows directly from Theorem 3.68 in [50], which is heavily based on Theorems 7.1 and 7.3 from [4].

6.2 Numerical examples for 2D photonic crystals

In this section we want to test the Fourier-Galerkin discretization with the benchmark problems, also considered in [16, 17, 21, 55, 60, 62], of computing band structures of 2D photonic crystals consisting of quadratic and circular rods.

6.2.1 Quadratic rods

First we consider structures with quadratic rods as depicted in Figure 6.2. This means that the two-valued functions $\varepsilon : \mathbb{R}^2 \to \mathbb{R}$ which describe the photonic crystal in our equations are of the form

$$\varepsilon(\boldsymbol{x}) = \begin{cases} a & \text{if } \|\boldsymbol{x}\|_\infty \le r, \\ 1 & \text{else.} \end{cases}$$

Now we compute the Fourier coefficients $\widehat{\varepsilon}_{\boldsymbol{n}}$ of the crystal function ε. We start with $\widehat{\varepsilon}_{\boldsymbol{0}}$:

$$\widehat{\varepsilon}_{\binom{0}{0}} \;=\; 1 + (a-1)(2r)^2 = 1 + 4(a-1)r^2.$$

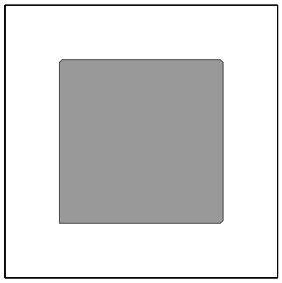


Figure 6.2: Example of a two-dimensional photonic crystal with a quadratic rod structure and rod width $2r$, where $0 < r < 0.5$, which attains the value a in the dark region and the value 1 in the white region.

Next we consider the case when for the index $\boldsymbol{n}$ holds $n_1 \neq 0$ and $n_2 \neq 0$:

$$\begin{aligned}
\widehat{\varepsilon}_{\binom{n_1}{n_2}} &= (a-1)\int_{-r}^{r}\int_{-r}^{r} e^{-\mathrm{i}2\pi(n_1x_1+n_2x_2)}dx_1dx_2 \\
&= (a-1)\left[\frac{\mathrm{i}}{2\pi n_1}e^{-\mathrm{i}2\pi n_1x_1}\right]_{-r}^{r}\left[\frac{\mathrm{i}}{2\pi n_2}e^{-\mathrm{i}2\pi n_2x_2}\right]_{-r}^{r} \\
&= -\frac{(a-1)}{4\pi^2 n_1n_2}\left(e^{-\mathrm{i}2\pi n_1r}-e^{\mathrm{i}2\pi n_1r}\right)\left(e^{-\mathrm{i}2\pi n_2r}-e^{\mathrm{i}2\pi n_2r}\right) \\
&= -\frac{(a-1)}{4\pi^2 n_1n_2}(-2\mathrm{i}\sin(2\pi n_1r))(-2\mathrm{i}\sin(2\pi n_2r)) \\
&= \frac{(a-1)}{\pi^2 n_1n_2}\sin(2\pi n_1r)\sin(2\pi n_2r)
\end{aligned}$$

Finally the case $n_1 \neq 0$, $n_2 = 0$

$$\begin{aligned}
\widehat{\varepsilon}_{\binom{n_1}{0}} &= 2r(a-1)\int_{-r}^{r} e^{-\mathrm{i}2\pi n_1x_1}dx_1 \\
&= \frac{\mathrm{i}2r(a-1)}{2\pi n_1}(-2\mathrm{i}\sin(2\pi n_1r)) \\
&= \frac{2r(a-1)}{\pi n_1}\sin(2\pi n_1r)
\end{aligned}$$

and $n_1 = 0$, $n_2 \neq 0$

$$\widehat{\varepsilon}_{\binom{0}{n_2}} = 2r(a-1)\int_{-r}^{r} e^{-\mathrm{i}2\pi n_2x_2}dx_2$$

$$\begin{aligned} &= \frac{\mathrm{i}2r(a-1)}{2\pi n_2}(-2\mathrm{i}\sin(2\pi n_2 r)) \\ &= \frac{2r(a-1)}{\pi n_2}\sin(2\pi n_2 r). \end{aligned}$$

We obtain

$$\widehat{\varepsilon}_{\boldsymbol{n}} = \begin{cases} 1+4(a-1)r^2 & \text{for } n_1=0,\ n_2=0, \\ \dfrac{(a-1)}{\pi^2 n_1 n_2}\sin(2\pi n_1 r)\sin(2\pi n_2 r) & \text{for } n_1\neq 0,\ n_2\neq 0, \\ \dfrac{2r(a-1)}{\pi n_1}\sin(2\pi n_1 r) & \text{for } n_1\neq 0,\ n_2=0, \\ \dfrac{2r(a-1)}{\pi n_2}\sin(2\pi n_2 r) & \text{for } n_1=0,\ n_2\neq 0. \end{cases} \tag{6.27}$$

6.2.2 Circular rods

Next we consider structures with circular rods as depicted in Figure 6.2. This means that the two-valued functions $\varepsilon : \mathbb{R}^2 \to \mathbb{R}$ which describe the photonic crystal in our equations are of the form

$$\varepsilon(\boldsymbol{x}) = \begin{cases} a & \text{if } \|\boldsymbol{x}\| \leq r, \\ 1 & \text{else.} \end{cases}$$

First we compute the Fourier coefficients $\widehat{\varepsilon}_{\boldsymbol{n}}$ of the crystal function ε. We start with $\widehat{\varepsilon}_{\boldsymbol{0}}$:

$$\widehat{\varepsilon}_{\binom{0}{0}} = 1+(a-1)\pi r^2.$$

For the representation of the coefficients $\widehat{\varepsilon}_{\boldsymbol{n}}$ with $\boldsymbol{n} \neq \boldsymbol{0}$ we need the *Bessel functions* J_k, which are defined as

$$\mathrm{J}_k(x) := \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-\mathrm{i}(k\tau - x\sin(\tau))}d\tau$$

for $k \in \mathbb{N}_0$. Using the identity

$$\mathrm{J}_0(2\pi\,\|\boldsymbol{n}\|\,\nu) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-\mathrm{i}(0(\pi+\theta)+2\pi\|\boldsymbol{n}\|\nu\sin(\pi+\theta))}d\theta$$

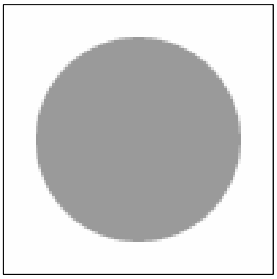


Figure 6.3: Example of a two-dimensional photonic crystal with a circular rod structure and rod radius $0 < r < 1/2$, which attains the value a in the dark region and the value 1 in the white region.

we compute the Fourier coefficients with indices $\boldsymbol{n} \neq \mathbf{0}$:

$$\begin{aligned}
\widehat{\varepsilon}_{\binom{n_1}{n_2}} &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} \varepsilon(\boldsymbol{x}) e^{-\mathrm{i}2\pi(n_1 x_1 + n_2 x_2)} dx_1 dx_2 \\
&= (a-1) \int_0^r \int_{-\pi}^{\pi} \left(\cos(2\pi \|\boldsymbol{n}\| \nu \cos(\theta)) - \mathrm{i}\sin(2\pi \|\boldsymbol{n}\| \nu \cos(\theta))\right) \nu d\theta d\nu \\
&= (a-1) \int_0^r \nu \mathrm{J}_0(2\pi \|\boldsymbol{n}\| \nu) d\nu \\
&= \frac{(a-1)r}{\|\boldsymbol{n}\|} \mathrm{J}_1(2\pi \|\boldsymbol{n}\| r).
\end{aligned}$$

We obtain

$$\widehat{\varepsilon}_{\boldsymbol{n}} = \begin{cases} 1 + (a-1)\pi r^2 & \text{for } (n_1, n_2) = (0,0), \\ \dfrac{(a-1)r}{\|\boldsymbol{n}\|} \mathrm{J}_1(2\pi \|\boldsymbol{n}\| r) & \text{for } (n_1, n_2) \neq (0,0). \end{cases} \tag{6.28}$$

As we have seen we do not need to assemble Toeplitz matrices whose entries are Fourier coefficients of ε, but rather vectors. The reason for that was that we can do fast Toeplitz products as a pointwise product in Fourier space. In our implementation for band structure computations of structures as discussed above we will use the formulas for the Fourier coefficients that we just have derived.

6.2.3 Numerical examples band structure

First we consider an example for a two-dimensional photonic crystal that consists of circular rods. We consider the same example as in [21, 55], namely a crystal that can be described by the crystal function

$$\varepsilon(\boldsymbol{x}) = \begin{cases} 9 & \text{if } \|\boldsymbol{x}\| \leq 0.38, \\ 1 & \text{else.} \end{cases}$$

In Figure 6.4 we see the numerical results for the Laplace-type equation and the divergence-type equation. The second example that we consider is an example for a two-dimensional photonic crystal that consists of quadratic rods. We consider the same example as in [21, 55], namely a crystal that can described by the crystal function

$$\varepsilon(\boldsymbol{x}) = \begin{cases} 9 & \text{if } \|\boldsymbol{x}\|_\infty \leq 0.3, \\ 1 & \text{else.} \end{cases}$$

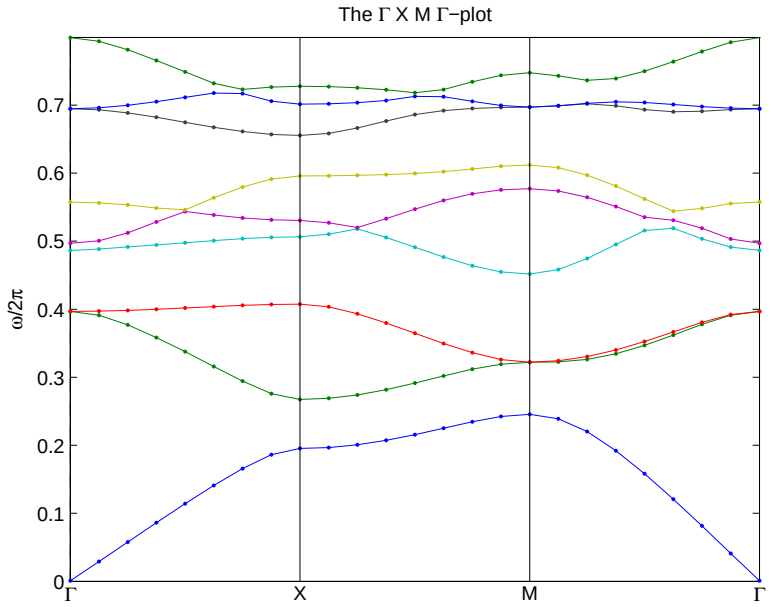
In Figure 6.5 we see the numerical results for the Laplace-type equation and the divergence-type equation. Both examples were computed with the Fourier-Galerkin method for $N = 255$, which corresponds to 261121 unknowns. It is interesting, that for the Laplace-type problem we were able to write the discrete equation as

$$D^{-1}(\boldsymbol{k}) [\![\varepsilon]\!] \widehat{\boldsymbol{u}} = \frac{1}{\lambda_N} \widehat{\boldsymbol{u}},$$

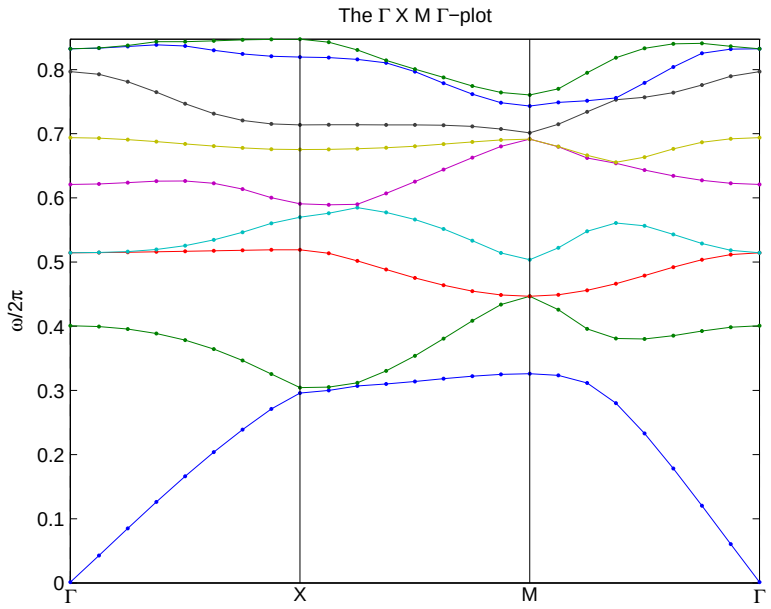
for a few largest eigenvalues. This allows us to use the `eigs` package in MATLAB® very efficiently. Since we are interested in largest eigenvalues of a matrix we only need to hand over to `eigs` a function handle for matrix-vector products. For one fixed $\boldsymbol{k}$ the Laplace-type problem in Figure 6.5 was solved in roughly 4 seconds on a desktop PC. In contrast to this, for the divergence-type problem we obtained the equation

$$C\widehat{\boldsymbol{u}} = \lambda \widehat{\boldsymbol{u}},$$

for the smallest eigenvalues. This means that we have to hand over to `eigs` a function handle for the solution of linear systems. This explains why for one fixed $\boldsymbol{k}$ the divergence-type problem in Figure 6.5 now was solved in roughly 140 seconds. Even much finer discretizations can be chosen on a desktop PC. For one fixed $\boldsymbol{k}$ and $N = 511$, which corresponds to 1046529 unknowns, the Laplace-type problem was solved in roughly 16 seconds and the divergence-type problem in roughly 560 seconds. Notice that we have have a factor of 4 between 140 and 560. This comes from the fact that we apply the FFT to a data set which is roughly 4 times larger.

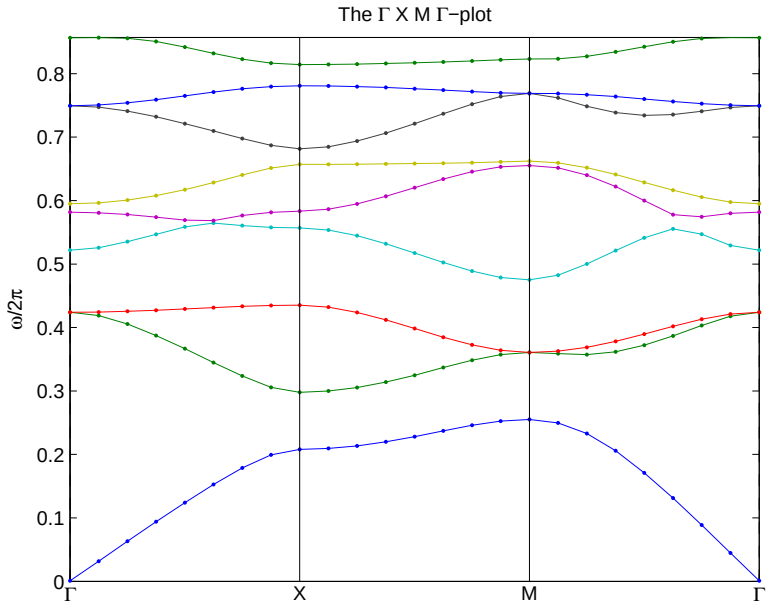


(a)

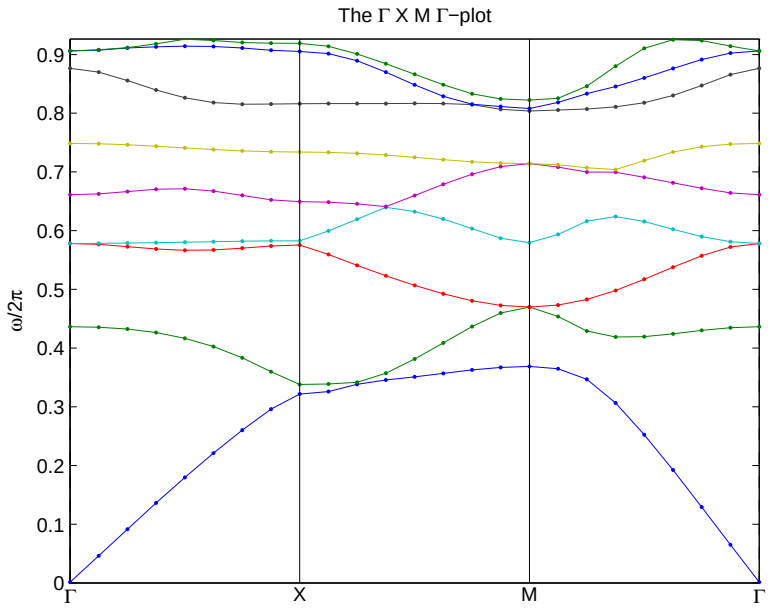


(b)

Figure 6.4: Band structure for a photonic crystal consisting of circular rods. (a) Laplace-type equation. (b) Divergence-type equation.



(a)



(b)

Figure 6.5: Band structure for a photonic crystal consisting of quadratic rods. (a) Laplace-type equation. (b) Divergence-type equation.

6.2.4 Experimental order of convergence

Now we want to verify numerically the convergence result from Section 6.1.4. We choose the example from the previous subsection with quadratic rod structure, i.e. the crystal function ε is given by

$$\varepsilon(\boldsymbol{x}) = \begin{cases} 9 & \text{if } \|\boldsymbol{x}\|_\infty \leq 0.3, \\ 1 & \text{else.} \end{cases}$$

In our test configuration we choose the parameter $\boldsymbol{k}$ as

$$\boldsymbol{k} = \begin{pmatrix} \pi \\ 0 \end{pmatrix}.$$

We compute a reference solution, and measure the errors of the discretizations against the computed reference solution. Our reference solution was computed with truncation order $N = 1984$, i.e. with 15752961 unknowns. In Figure 6.6 we see the eigenfunction error in the H^1_{per} norm, and in Figure 6.7 we see the error for the eigenvalues. We compute the maximum over all errors for the first five eigenvalues and corresponding eigenfunctions.

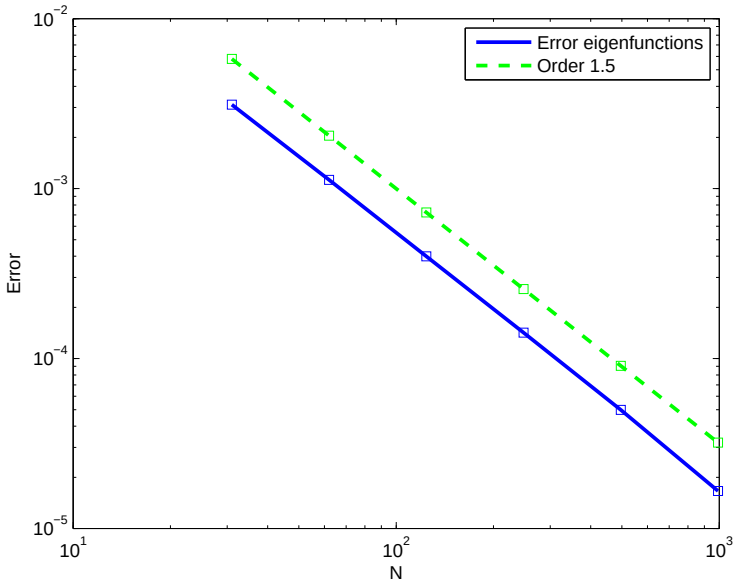


Figure 6.6: Experimental order of convergence for the eigenfunctions.

For the divergence-type equation we have observed experimental order of convergence 0.5 for the eigenfunctions and 1 for the eigenvalues. This suggests that the eigenfunctions are in $H^{3/2-\rho}_{\text{per}}(\Omega)$.

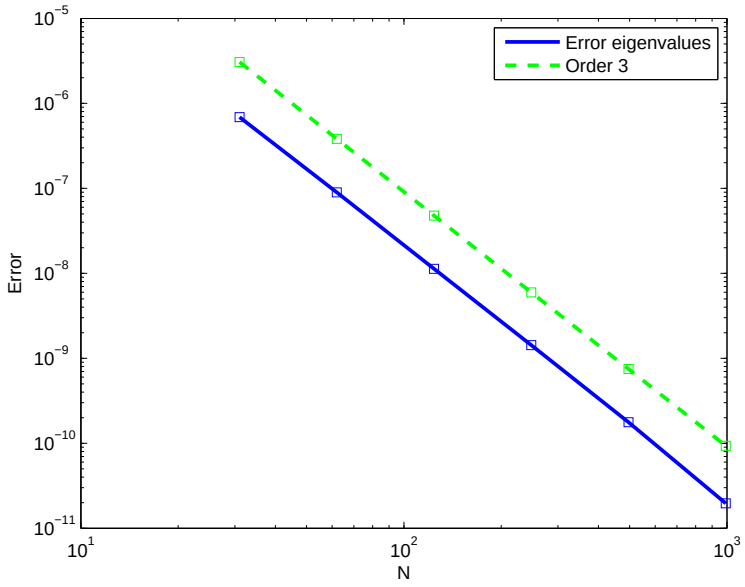


Figure 6.7: Experimental order of convergence for the eigenvalues.

6.2.5 Remark to inexact coefficients

We have discussed the Fourier-Galerkin method for a situation, where we have access to the exact Fourier coefficients of the crystal function ε. However, for crystal functions ε that represent more complicated geometries we will in general not be able to find explicit representations for its Fourier coefficients. In practice it is common to use the FFT of the nodal values of ε on an equidistant grid in order to approximate the Fourier coefficients of ε. Due to the error that is being made in this process we should choose the parameter N_g (number of grid points per dimension) for the grid larger than the truncation order N for the Fourier-Galerkin method. Notice that we only need to apply this FFT once in the beginning, in order to compute the approximate Fourier coefficients. Moreover, since for the matrix vector products, that we permanently use for the iterative methods, we actually work with FFTs of twice the dimension, it is actually the same cost to choose $N_g \approx 2(2N+1)$ as for one matrix-vector product in the iterations. Choosing $N_g \approx 4(2N+1)$ is approximately as time consuming as 2 matrix vector products that we perform for our iterations, and choosing $N_g \approx 8(2N+1)$ takes about the time of 8 matrix vector products. In [51] the question about additional errors, which are introduced by sampling of the coefficients via FFT, is discussed. This question is discussed for a Schrödinger operator. It was shown that in addition to the error made in the Galerkin approximation the error introduced by sampling the

coefficients behaves like $\mathcal{O}\left(N_g^{-1/2+\varepsilon}\right)$. However, they also commented that this result is not sharp and in numerical tests rather an additional error of $\mathcal{O}(N_g^{-1})$ was observed. Since the Helmholtz problem, that we have considered here, is quite similar to the problem considered in [51] we expect a similar behaviour.

6.3 3D Helmholtz equation

In this section we consider the Helmholtz equation equation in three dimensions:

$$-\Delta u(\boldsymbol{x}) = \omega^2 \varepsilon(\boldsymbol{x}) u(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \mathbb{R}^3, \tag{6.29}$$

where ε is $\mathbb{Z}^3$-periodic, piecewise smooth and and satisfies $0 < \varepsilon_{\min} \leq \varepsilon(\boldsymbol{x}) \leq \varepsilon_{\max} < \infty$. Although solving this problem numerically does not correspond to a photonic band strutcure calculation as the previous problems, we want to discuss how to solve this eigenvalue problem numerically, because it is very similar to the two-dimensional case. Exactly the same way as in Section 6.1.3 we can discretize the problem. If we follow the same same steps we obtain a generalized matrix eigenvalue problem

$$D(\boldsymbol{k})\widehat{\boldsymbol{u}} = \lambda_N [\![\varepsilon]\!] \widehat{\boldsymbol{u}},$$

where $[\![\varepsilon]\!]$ is a Block-Toeplitz-Toeplitz-Block matrix of level three generated by the Fourier coefficients of ε and $D(\boldsymbol{k})$ is a diagonal matrix depending on the parameter $\boldsymbol{k} \in B \setminus \{\mathbf{0}\}$. Due to the lexicographic ordering of our indices we obtain the following structure for the matrix $[\![\varepsilon]\!]$ (with the four Toeplitz blocks in the corners)

$$[\![\varepsilon]\!] := \begin{pmatrix}
\widehat{\varepsilon}_{\binom{0}{0}\binom{0}{0}\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{0}\binom{0}{0}\binom{0}{0}} & & & & \widehat{\varepsilon}_{\binom{0}{-2N}\binom{0}{-2N}\binom{0}{-2N}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{-2N}\binom{0}{-2N}\binom{0}{-2N}} \\
\vdots & \ddots & \vdots & \cdots & \cdots & \cdots & \vdots & \ddots & \vdots \\
\widehat{\varepsilon}_{\binom{2N}{0}\binom{0}{0}\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{0}{0}\binom{0}{0}\binom{0}{0}} & & & & \widehat{\varepsilon}_{\binom{2N}{-2N}\binom{0}{-2N}\binom{0}{-2N}} & \cdots & \widehat{\varepsilon}_{\binom{0}{-2N}\binom{0}{-2N}\binom{0}{-2N}} \\
& \vdots & & \ddots & & & & \vdots & \\
& \vdots & & & \ddots & & & \vdots & \\
& \vdots & & & & \ddots & & \vdots & \\
\widehat{\varepsilon}_{\binom{0}{2N}\binom{0}{0}\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{2N}\binom{0}{0}\binom{0}{0}} & & & & \widehat{\varepsilon}_{\binom{0}{0}\binom{0}{0}\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{-2N}{0}\binom{0}{0}\binom{0}{0}} \\
\vdots & \ddots & \vdots & \cdots & \cdots & \cdots & \vdots & \ddots & \vdots \\
\widehat{\varepsilon}_{\binom{2N}{2N}\binom{0}{0}\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{0}{2N}\binom{0}{0}\binom{0}{0}} & & & & \widehat{\varepsilon}_{\binom{2N}{0}\binom{0}{0}\binom{0}{0}} & \cdots & \widehat{\varepsilon}_{\binom{0}{0}\binom{0}{0}\binom{0}{0}}
\end{pmatrix},$$

and the following structure for the diagonal matrix

$$D(\boldsymbol{k}) := \begin{pmatrix} \left|2\pi\begin{pmatrix}-N\\-N\\-N\end{pmatrix}+\boldsymbol{k}\right|^2 & & & & & & \\ & \ddots & & & & & \\ & & \left|2\pi\begin{pmatrix}N\\-N\\-N\end{pmatrix}+\boldsymbol{k}\right|^2 & & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \\ & & & & & \left|2\pi\begin{pmatrix}-N\\N\\N\end{pmatrix}+\boldsymbol{k}\right|^2 & \\ & & & & & & \ddots \\ & & & & & & & \left|2\pi\begin{pmatrix}N\\N\\N\end{pmatrix}+\boldsymbol{k}\right|^2 \end{pmatrix}.$$

Notice that $D(\boldsymbol{k})$ is invertible for $\boldsymbol{k} \neq \boldsymbol{0}$. Again we can consider the equation

$$D^{-1}(\boldsymbol{k})\,[\![\varepsilon]\!]\,\widehat{\boldsymbol{u}} = \frac{1}{\lambda_N}\widehat{\boldsymbol{u}},$$

in order to compute the largest eigenvalues. This can be done very efficiently by applying fast Toeplitz products in iterative methods as we have seen.

Chapter 7

3D Maxwell Problem

In this chapter we consider the problem of discretizing the 3D Maxwell equations with the Fourier-Galerkin method. First we consider the H-field formulation, then the E-field formulation. Through the whole chapter ε will denote a $\mathbb{Z}^3$-periodic piecewise constant function which satisfies $0 < \varepsilon_{\min} \leq \varepsilon(\boldsymbol{x}) \leq \varepsilon_{\max} < \infty$. The unit cell will be $\Omega := (-\frac{1}{2}, \frac{1}{2})^3$ and the index range is defined as

$$\mathbb{I}_N := \left\{ \boldsymbol{n} \in \mathbb{Z}^3 \,:\, \|\boldsymbol{n}\|_\infty \leq N \right\}.$$

Similarly as in [35, 39] we will decompose the vectorial 3D problem into three scalar 3D components, since it is desirable to apply the same techniques as for the Helmholtz problem. We have seen that reducing the computation of eigenvalues to a block Toeplitz product computation can be done very efficiently, concerning computational and memory costs, via the FFT. For the discretizations we will consider the parametrized problems that were introduced in Section 3.4. Now let $\boldsymbol{k} \in B$ be given and $\omega \neq 0$. It can be shown, that if $(\lambda, \boldsymbol{h})$ is an eigensolution of (3.13), then $(\lambda, \boldsymbol{e})$ with

$$\boldsymbol{e} = -\frac{\mathrm{i}}{\omega\varepsilon} (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{h}$$

is an eigensolution to (3.15). Similarly, it can be shown that if $(\lambda, \boldsymbol{e})$ is an eigensolution of (3.15), then $(\lambda, \boldsymbol{h})$ with

$$\boldsymbol{h} = \frac{\mathrm{i}}{\omega} (\nabla + \mathrm{i}\boldsymbol{k}) \times \boldsymbol{e}$$

is an eigensolution to (3.13). This is the reason why it is sufficient to solve either the H-field formulation or the E-field formulation. If we are interested in the band structure only, then it does not matter which one we choose to solve because the eigenvalues are the same. In the following we discuss the discretizations to both problems.

7.1 H-field formulation

The H-field formulation reads

$$(\nabla + \mathrm{i}\boldsymbol{k})\times \left(\frac{1}{\varepsilon} (\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{h} \right) = \omega^2 \boldsymbol{h} \quad \text{in } \Omega, \tag{7.1}$$

$$(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{h} = 0 \quad \text{in } \Omega. \tag{7.2}$$

The weak form is to find $\lambda \in \mathbb{R}$ and $\mathbf{0} \neq \boldsymbol{u} \in \boldsymbol{H}^1_p(\Omega)$ such that

$$a_{\boldsymbol{k}}(\boldsymbol{u}, \boldsymbol{v}) = \lambda b(\boldsymbol{u}, \boldsymbol{v})$$

holds for every $\boldsymbol{v} \in \boldsymbol{H}^1_p(\Omega)$, where

$$a_{\boldsymbol{k}}(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \frac{1}{\varepsilon(\boldsymbol{x})} (\nabla + \mathrm{i}\boldsymbol{k})\boldsymbol{u}(\boldsymbol{x}) \cdot \overline{(\nabla + \mathrm{i}\boldsymbol{k})\boldsymbol{v}(\boldsymbol{x})}\, d\boldsymbol{x}$$

and

$$b(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega \boldsymbol{u}(\boldsymbol{x}) \cdot \overline{\boldsymbol{v}(\boldsymbol{x})}\, d\boldsymbol{x}.$$

In order to apply the Fourier-Galerkin method we choose a finite dimensional subspace of the Sobolev space of periodic functions $\boldsymbol{H}^1_p(\Omega)$, which is the span of a finite number of plane waves:

$$\mathcal{T}_N := \operatorname{span} \left\{ e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \,:\, \|\boldsymbol{n}\|_\infty \leq N \right\}^3.$$

This means that any element $\boldsymbol{u}_N \in \mathcal{T}_N$ can be represented as

$$\boldsymbol{u}_N(\boldsymbol{x}) = \sum_{\boldsymbol{n}\in\mathbb{I}_N} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \widehat{\boldsymbol{u}}_{\boldsymbol{n}} = \sum_{\boldsymbol{n}\in\mathbb{I}_N} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \begin{pmatrix} \widehat{u}^{(1)}_{\boldsymbol{n}} \\ \widehat{u}^{(2)}_{\boldsymbol{n}} \\ \widehat{u}^{(3)}_{\boldsymbol{n}} \end{pmatrix} = \sum_{\boldsymbol{n}\in\mathbb{I}_N} \begin{pmatrix} \widehat{u}^{(1)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}^{(2)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}^{(3)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \end{pmatrix}.$$

Applying the Galerkin method means that we have to find $\lambda_N \in \mathbb{R}$ and $\mathbf{0} \neq \boldsymbol{u}_N \in \mathcal{T}_N$ such that

$$a_{\boldsymbol{k}}(\boldsymbol{u}_N, \boldsymbol{v}_N) = \lambda_N b(\boldsymbol{u}_N, \boldsymbol{v}_N) \quad \text{for all } \boldsymbol{v}_N \in \mathcal{T}_N$$

holds. We consider

$$a_{\boldsymbol{k}}(\boldsymbol{u}_N, \boldsymbol{u}^{(j)}_{\boldsymbol{m}}) = \lambda_N b(\boldsymbol{u}_N, \boldsymbol{u}^{(j)}_{\boldsymbol{m}}) \tag{7.3}$$

for $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$, with the test functions

$$\boldsymbol{u}_{\boldsymbol{m}}^{(1)}(\boldsymbol{x}) = \begin{pmatrix} e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}} \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{u}_{\boldsymbol{m}}^{(2)}(\boldsymbol{x}) = \begin{pmatrix} 0 \\ e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}} \\ 0 \end{pmatrix}, \quad \boldsymbol{u}_{\boldsymbol{m}}^{(3)}(\boldsymbol{x}) = \begin{pmatrix} 0 \\ 0 \\ e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}} \end{pmatrix}. \quad (7.4)$$

Alternatively we can also represent the test functions as $\boldsymbol{u}_{\boldsymbol{m}}^{(j)} = e^{\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\boldsymbol{e}_j$, for $j = 1, 2, 3$ and $\boldsymbol{m} \in \mathbb{I}_N$, where $\boldsymbol{e}_j \in \mathbb{R}^3$ is the j-th unit vector. With

$$(\nabla + \mathrm{i}\boldsymbol{k})\times \sum_{\boldsymbol{n}\in\mathbb{I}_N} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\widehat{\boldsymbol{u}}_{\boldsymbol{n}} = \mathrm{i}\sum_{\boldsymbol{n}\in\mathbb{I}_N} (2\pi\boldsymbol{n} + \boldsymbol{k}) \times \widehat{\boldsymbol{u}}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}$$

and

$$(\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{u}_{\boldsymbol{m}}^{(j)}(\boldsymbol{x}) = -\mathrm{i}(2\pi\boldsymbol{m} + \boldsymbol{k}) \times \boldsymbol{e}_j\, e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}$$

we obtain

$$\begin{aligned} a_{\boldsymbol{k}}(\boldsymbol{u}_N, \boldsymbol{u}_{\boldsymbol{m}}^{(j)}) &= \int_\Omega \frac{1}{\varepsilon(\boldsymbol{x})} \left((\nabla + \mathrm{i}\boldsymbol{k})\times \sum_{\boldsymbol{n}\in\mathbb{I}_N} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\widehat{\boldsymbol{u}}_{\boldsymbol{n}} \right) \cdot \overline{\left((\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{u}_{\boldsymbol{m}}^{(j)}(\boldsymbol{x}) \right)}\, d\boldsymbol{x} \\ &= \sum_{\boldsymbol{n}\in\mathbb{I}_N} (2\pi\boldsymbol{n} + \boldsymbol{k}) \times \widehat{\boldsymbol{u}}_{\boldsymbol{n}} \cdot (2\pi\boldsymbol{m} + \boldsymbol{k}) \times \boldsymbol{e}_j \int_\Omega \frac{1}{\varepsilon(\boldsymbol{x})} e^{-\mathrm{i}2\pi(\boldsymbol{m}-\boldsymbol{n})\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\left(\frac{1}{\varepsilon}\right)}_{\boldsymbol{m}-\boldsymbol{n}} (2\pi\boldsymbol{n} + \boldsymbol{k}) \times \widehat{\boldsymbol{u}}_{\boldsymbol{n}} \cdot (2\pi\boldsymbol{m} + \boldsymbol{k}) \times \boldsymbol{e}_j \qquad (7.5) \end{aligned}$$

for all $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$. Moreover, we obtain

$$\begin{aligned} b(\boldsymbol{u}_N, \boldsymbol{u}_{\boldsymbol{m}}^{(j)}) &= \int_\Omega \left(\sum_{\boldsymbol{n}\in\mathbb{I}_N} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\widehat{\boldsymbol{u}}_{\boldsymbol{n}} \right) \cdot \overline{\boldsymbol{u}_{\boldsymbol{m}}^{(j)}(\boldsymbol{x})}\, d\boldsymbol{x} \\ &= \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\boldsymbol{u}}_{\boldsymbol{n}} \cdot \boldsymbol{e}_j \int_\Omega e^{\mathrm{i}2\pi(\boldsymbol{n}-\boldsymbol{m})\cdot\boldsymbol{x}}\, d\boldsymbol{x} \\ &= \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\boldsymbol{u}}_{\boldsymbol{n}}^{(j)} \delta_{\boldsymbol{m},\boldsymbol{n}} \\ &= \widehat{\boldsymbol{u}}_{\boldsymbol{m}}^{(j)} \qquad (7.6) \end{aligned}$$

for all $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$. This means that the discretization via (7.3) leads to a usual matrix eigenvalue problem. Next we consider what kind of structure the matrix has, that we obtain from considering (7.5) for all $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$.

We first count $\boldsymbol{m} \in \mathbb{I}_N$ for $j = 1$, then for $j = 2$ and finally for $j = 3$. This means we choose the ordering for the unknown coefficients by starting with all the first components, then all the second components and finally all the third components. From (7.5) and (7.6) it follows, that we consider an ordinary matrix eigenvalue problem for the smallest eigenvalues. The Hermitian eigenvalue problem reads

$$B^\top A^\top T\, A\, B\widehat{\boldsymbol{u}} = \lambda\widehat{\boldsymbol{u}}, \tag{7.7}$$

where A is of the form

$$A = \left(\begin{array}{c|c|c} 0 & -A_3 & A_2 \\ \hline A_3 & 0 & -A_1 \\ \hline -A_2 & A_1 & 0 \end{array}\right),$$

with diagonal matrices A_1, A_2 and A_3 that contain the entries which arise from the cross product of vectors in (7.5) according to (2.2), B is the permutation matrix which sorts the vector with the Fourier coefficients componentwise as explained above and T is a Hermitian block-diagonal matrix with the Toeplitz blocks $[\![\frac{1}{\varepsilon}]\!]$, i.e.

$$T = \begin{pmatrix} [\![\frac{1}{\varepsilon}]\!] & & \\ & [\![\frac{1}{\varepsilon}]\!] & \\ & & [\![\frac{1}{\varepsilon}]\!] \end{pmatrix}.$$

Notice that for $d := (2N+1)^3$ we have $A_1, A_2, A_3 \in \mathbb{R}^{d\times d}$ and $A, B, T \in \mathbb{R}^{3d\times 3d}$. This means that we can, similarly as we did for the scalar problems, decompose the matrix which results from the discretization into a product of sparse matrices and BTTB matrices. This decomposition can be used in order to compute matrix-vector products efficiently when an iterative eigensolver is applied.

7.2 E-field formulation

The E-field formulation reads

$$(\nabla + \mathrm{i}\boldsymbol{k})\times(\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{e} = \omega^2\varepsilon\boldsymbol{e} \quad \text{in } \Omega, \tag{7.8}$$

$$(\nabla + \mathrm{i}\boldsymbol{k})\cdot(\varepsilon\boldsymbol{e}) = 0 \quad \text{in } \Omega. \tag{7.9}$$

The weak form is to find $\lambda \in \mathbb{R}$ and $\mathbf{0} \neq \boldsymbol{u} \in \boldsymbol{H}^1_p(\Omega)$ such that

$$a_{\boldsymbol{k}}(\boldsymbol{u}, \boldsymbol{v}) = \lambda b(\boldsymbol{u}, \boldsymbol{v})$$

holds for every $\boldsymbol{v} \in \boldsymbol{H}^1_p(\Omega)$, where

$$a_{\boldsymbol{k}}(\boldsymbol{u}, \boldsymbol{v}) := \int_\Omega ((\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{u}) \cdot \overline{((\nabla + \mathrm{i}\boldsymbol{k})\times \boldsymbol{v})}\, d\boldsymbol{x}$$

and

$$b(\boldsymbol{u},\boldsymbol{v}) := \int_\Omega \varepsilon\, \boldsymbol{u}\cdot\overline{\boldsymbol{v}}\, d\boldsymbol{x}.$$

Applying the Galerkin method means that we have to find $\lambda_N \in \mathbb{R}$ and $\mathbf{0} \neq \boldsymbol{u}_N \in \mathcal{T}_N$ such that

$$a_{\boldsymbol{k}}(\boldsymbol{u}_N,\boldsymbol{v}_N) = \lambda_N b(\boldsymbol{u}_N,\boldsymbol{v}_N) \quad \text{for all } \boldsymbol{v}_N \in \mathcal{T}_N$$

holds. We consider

$$a_{\boldsymbol{k}}(\boldsymbol{u}_N,\boldsymbol{u}_{\boldsymbol{n}}^{(j)}) = \lambda_N b(\boldsymbol{u}_N,\boldsymbol{u}_{\boldsymbol{n}}^{(j)})$$

for $j=1,2,3$ and all $\boldsymbol{m}\in\mathbb{I}_N$, where the test functions are the same as in (7.4). This means we consider

$$\begin{aligned}\int_\Omega \left((\nabla+\mathrm{i}\boldsymbol{k})\times\left(\sum_{\boldsymbol{n}\in\mathbb{I}_\mathrm{N}}\begin{pmatrix}\widehat{u}_{\boldsymbol{n}}^{(1)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\\ \widehat{u}_{\boldsymbol{n}}^{(2)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\\ \widehat{u}_{\boldsymbol{n}}^{(3)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\end{pmatrix}\right)\right)\cdot\overline{\left((\nabla+\mathrm{i}\boldsymbol{k})\times\boldsymbol{u}_{\boldsymbol{n}}^{(j)}(\boldsymbol{x})\right)}\,d\boldsymbol{x}\\ = \lambda_N\int_\Omega\left(\varepsilon(\boldsymbol{x})\sum_{\boldsymbol{n}\in\mathbb{I}_\mathrm{N}}\begin{pmatrix}\widehat{u}_{\boldsymbol{n}}^{(1)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\\ \widehat{u}_{\boldsymbol{n}}^{(2)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\\ \widehat{u}_{\boldsymbol{n}}^{(3)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\end{pmatrix}\right)\cdot\overline{\boldsymbol{u}_{\boldsymbol{n}}^{(j)}(\boldsymbol{x})}\,d\boldsymbol{x},\end{aligned}$$

for $j=1,2,3$ and all $\boldsymbol{m}\in\mathbb{I}_N$. Now we consider

$$\begin{aligned}(\nabla+\mathrm{i}\boldsymbol{k})\times\begin{pmatrix}\widehat{u}_{\boldsymbol{n}}^{(1)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\\ \widehat{u}_{\boldsymbol{n}}^{(2)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\\ \widehat{u}_{\boldsymbol{n}}^{(3)}e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}\end{pmatrix} &= \mathrm{i}(2\pi\boldsymbol{n}+\mathbf{k})\times\begin{pmatrix}\widehat{u}_{\boldsymbol{n}}^{(1)}\\ \widehat{u}_{\boldsymbol{n}}^{(2)}\\ \widehat{u}_{\boldsymbol{n}}^{(3)}\end{pmatrix}e^{\mathrm{i}2\pi\mathbf{n}\cdot\boldsymbol{x}}\\ &= \mathrm{i}e^{\mathrm{i}2\pi\mathbf{n}\cdot\boldsymbol{x}}\begin{pmatrix}(2\pi n_2+k_2)\widehat{u}_{\boldsymbol{n}}^{(3)}-(2\pi n_3+k_3)\widehat{u}_{\boldsymbol{n}}^{(2)}\\ (2\pi n_3+k_3)\widehat{u}_{\boldsymbol{n}}^{(1)}-(2\pi n_1+k_1)\widehat{u}_{\boldsymbol{n}}^{(3)}\\ (2\pi n_1+k_1)\widehat{u}_{\boldsymbol{n}}^{(2)}-(2\pi n_2+k_2)\widehat{u}_{\boldsymbol{n}}^{(1)}\end{pmatrix},\end{aligned}$$

and

$$\begin{aligned}(\nabla-\mathrm{i}\mathbf{k})\times\begin{pmatrix}e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\\ 0\\ 0\end{pmatrix} &= (-\mathrm{i})e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\begin{pmatrix}0\\ 2\pi m_3+k_3\\ -(2\pi m_2+k_2)\end{pmatrix},\\ (\nabla-\mathrm{i}\mathbf{k})\times\begin{pmatrix}0\\ e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\\ 0\end{pmatrix} &= (-\mathrm{i})e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}}\begin{pmatrix}-(2\pi m_3+k_3)\\ 0\\ (2\pi m_1+k_1)\end{pmatrix},\end{aligned}$$

$$(\nabla - \mathrm{i}\mathbf{k}) \times \begin{pmatrix} 0 \\ 0 \\ e^{-\mathrm{i}2\pi \boldsymbol{m}\cdot\boldsymbol{x}} \end{pmatrix} = (-\mathrm{i})e^{-\mathrm{i}2\pi \boldsymbol{m}\cdot\boldsymbol{x}} \begin{pmatrix} 2\pi m_2 + k_2 \\ -(2\pi m_1 + k_1) \\ 0 \end{pmatrix}.$$

Firstly, this leads to

$$\left((\nabla + \mathrm{i}\boldsymbol{k}) \times \begin{pmatrix} \widehat{u}_{\boldsymbol{n}}^{(1)} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}_{\boldsymbol{n}}^{(2)} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}_{\boldsymbol{n}}^{(3)} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \end{pmatrix} \right) \cdot \left((\nabla - \mathrm{i}\mathbf{k}) \times \begin{pmatrix} e^{-\mathrm{i}2\pi \boldsymbol{m}\cdot\boldsymbol{x}} \\ 0 \\ 0 \end{pmatrix} \right)$$

$$= e^{\mathrm{i}2\pi(\mathbf{n}-\boldsymbol{m})\cdot\boldsymbol{x}} \begin{pmatrix} (2\pi n_2 + k_2)\widehat{u}_{\boldsymbol{n}}^{(3)} - (2\pi n_3 + k_3)\widehat{u}_{\boldsymbol{n}}^{(2)} \\ (2\pi n_3 + k_3)\widehat{u}_{\boldsymbol{n}}^{(1)} - (2\pi n_1 + k_1)\widehat{u}_{\boldsymbol{n}}^{(3)} \\ (2\pi n_1 + k_1)\widehat{u}_{\boldsymbol{n}}^{(2)} - (2\pi n_2 + k_2)\widehat{u}_{\boldsymbol{n}}^{(1)} \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 2\pi m_3 + k_3 \\ -(2\pi m_2 + k_2) \end{pmatrix}$$

$$= \Bigg(\underbrace{\left((2\pi n_3 + k_3)(2\pi m_3 + k_3) + (2\pi n_2 + k_2)(2\pi m_2 + k_2)\right)}_{=:a_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}^{(1,1)}} \widehat{u}_{\boldsymbol{n}}^{(1)}$$

$$\underbrace{-(2\pi n_1 + k_1)(2\pi m_2 + k_2)}_{=:a_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}^{(1,2)}} \widehat{u}_{\boldsymbol{n}}^{(2)} \underbrace{-(2\pi n_1 + k_1)(2\pi m_3 + k_3)}_{=:a_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}^{(1,3)}} \widehat{u}_{\boldsymbol{n}}^{(3)} \Bigg) e^{\mathrm{i}2\pi(\mathbf{n}-\boldsymbol{m})\cdot\boldsymbol{x}},$$

secondly to

$$\left((\nabla + \mathrm{i}\boldsymbol{k}) \times \begin{pmatrix} \widehat{u}_{\boldsymbol{n}}^{(1)} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}_{\boldsymbol{n}}^{(2)} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}_{\boldsymbol{n}}^{(3)} e^{\mathrm{i}2\pi \boldsymbol{n}\cdot\boldsymbol{x}} \end{pmatrix} \right) \cdot \left((\nabla - \mathrm{i}\mathbf{k}) \times \begin{pmatrix} 0 \\ e^{-\mathrm{i}2\pi \boldsymbol{m}\cdot\boldsymbol{x}} \\ 0 \end{pmatrix} \right)$$

$$= e^{\mathrm{i}2\pi(\mathbf{n}-\boldsymbol{m})\cdot\boldsymbol{x}} \begin{pmatrix} (2\pi n_2 + k_2)\widehat{u}_{\boldsymbol{n}}^{(3)} - (2\pi n_3 + k_3)\widehat{u}_{\boldsymbol{n}}^{(2)} \\ (2\pi n_3 + k_3)\widehat{u}_{\boldsymbol{n}}^{(1)} - (2\pi n_1 + k_1)\widehat{u}_{\boldsymbol{n}}^{(3)} \\ (2\pi n_1 + k_1)\widehat{u}_{\boldsymbol{n}}^{(2)} - (2\pi n_2 + k_2)\widehat{u}_{\boldsymbol{n}}^{(1)} \end{pmatrix} \cdot \begin{pmatrix} -(2\pi m_3 + k_3) \\ 0 \\ 2\pi m_1 + k_1 \end{pmatrix}$$

$$= \Bigg(\underbrace{-(2\pi n_2 + k_2)(2\pi m_1 + k_1)}_{=:a_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}^{(2,1)}} \widehat{u}_{\boldsymbol{n}}^{(1)}$$

$$+ \underbrace{\left((2\pi n_3 + k_3)(2\pi m_3 + k_3) + (2\pi n_1 + k_1)(2\pi m_1 + k_1)\right)}_{=:a_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}^{(2,2)}} \widehat{u}_{\boldsymbol{n}}^{(2)}$$

$$\underbrace{-(2\pi n_2 + k_2)(2\pi m_3 + k_3)}_{=:a^{(2,3)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}} \widehat{u}^{(3)}_{\boldsymbol{n}}\Bigg) e^{\mathrm{i}2\pi(\mathbf{n}-\boldsymbol{m})\cdot\boldsymbol{x}},$$

and thirdly to

$$\left((\nabla + \mathrm{i}\boldsymbol{k})\times \begin{pmatrix} \widehat{u}^{(1)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}^{(2)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \\ \widehat{u}^{(3)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}} \end{pmatrix}\right) \cdot \left((\nabla - \mathrm{i}\mathbf{k}) \times \begin{pmatrix} 0 \\ 0 \\ e^{-\mathrm{i}2\pi\boldsymbol{m}\cdot\boldsymbol{x}} \end{pmatrix}\right)$$

$$= e^{\mathrm{i}2\pi(\mathbf{n}-\boldsymbol{m})\cdot\boldsymbol{x}} \begin{pmatrix} (2\pi n_2 + k_2)\widehat{u}^{(3)}_{\boldsymbol{n}} - (2\pi n_3 + k_3)\widehat{u}^{(2)}_{\boldsymbol{n}} \\ (2\pi n_3 + k_3)\widehat{u}^{(1)}_{\boldsymbol{n}} - (2\pi n_1 + k_1)\widehat{u}^{(3)}_{\boldsymbol{n}} \\ (2\pi n_1 + k_1)\widehat{u}^{(2)}_{\boldsymbol{n}} - (2\pi n_2 + k_2)\widehat{u}^{(1)}_{\boldsymbol{n}} \end{pmatrix} \cdot \begin{pmatrix} 2\pi m_2 + k_2 \\ -(2\pi m_1 + k_1) \\ 0 \end{pmatrix}$$

$$= \Bigg(\underbrace{-(2\pi n_3 + k_3)(2\pi m_1 + k_1)}_{=:a^{(3,1)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}} \widehat{u}^{(1)}_{\boldsymbol{n}} \underbrace{-(2\pi n_3 + k_3)(2\pi m_2 + k_2)}_{=:a^{(3,2)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}} \widehat{u}^{(2)}_{\boldsymbol{n}}$$

$$+ \underbrace{((2\pi n_2 + k_2)(2\pi m_2 + k_2) + (2\pi n_1 + k_1)(2\pi m_1 + k_1))}_{=:a^{(3,3)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}}} \widehat{u}^{(3)}_{\boldsymbol{n}} \Bigg) e^{\mathrm{i}2\pi(\mathbf{n}-\boldsymbol{m})\cdot\boldsymbol{x}}.$$

With

$$a_{\boldsymbol{k}}(\boldsymbol{u}_N, \boldsymbol{u}^{(j)}_{\boldsymbol{m}}) = \lambda_N b(\boldsymbol{u}_N, \boldsymbol{u}^{(j)}_{\boldsymbol{m}})$$

$$\iff \quad \sum_{\boldsymbol{n}\in\mathbb{I}_N} a_{\boldsymbol{k}}(\widehat{\boldsymbol{u}}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, \boldsymbol{u}^{(j)}_{\boldsymbol{m}}) = \lambda_N \sum_{\boldsymbol{n}\in\mathbb{I}_N} b(\widehat{\boldsymbol{u}}_{\boldsymbol{n}} e^{\mathrm{i}2\pi\boldsymbol{n}\cdot\boldsymbol{x}}, \boldsymbol{u}^{(j)}_{\boldsymbol{m}})$$

for $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$, we obtain

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} \int_\Omega \left(a^{(j,1)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}} \widehat{u}^{(1)}_{\boldsymbol{n}} + a^{(j,2)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}} \widehat{u}^{(2)}_{\boldsymbol{n}} + a^{(j,3)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}} \widehat{u}^{(3)}_{\boldsymbol{n}} \right) e^{\mathrm{i}2\pi(\boldsymbol{n}-\boldsymbol{m})\cdot\boldsymbol{x}}\, d\boldsymbol{x}$$

$$= \lambda_N \sum_{\boldsymbol{n}\in\mathbb{I}_N} \int_\Omega \varepsilon(\boldsymbol{x}) \widehat{\boldsymbol{u}}^{(j)}_{\boldsymbol{n}} e^{\mathrm{i}2\pi(\boldsymbol{n}-\boldsymbol{m})\cdot\boldsymbol{x}}\, d\boldsymbol{x}$$

for $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$. This is equivalent to

$$\sum_{\boldsymbol{n}\in\mathbb{I}_N} \left(a^{(j,1)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}} \widehat{u}^{(1)}_{\boldsymbol{n}} + a^{(j,2)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}} \widehat{u}^{(2)}_{\boldsymbol{n}} + a^{(j,3)}_{\boldsymbol{m},\boldsymbol{n},\boldsymbol{k}} \widehat{u}^{(3)}_{\boldsymbol{n}} \right) \delta_{\boldsymbol{m}\boldsymbol{n}} = \lambda_N \sum_{\boldsymbol{n}\in\mathbb{I}_N} \widehat{\varepsilon}_{\boldsymbol{m}-\boldsymbol{n}} \widehat{\boldsymbol{u}}^{(j)}_{\boldsymbol{n}}$$

for $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$, and finally leads to

$$a^{(j,1)}_{\boldsymbol{m},\boldsymbol{m},\boldsymbol{k}} \widehat{u}^{(1)}_{\boldsymbol{n}} + a^{(j,2)}_{\boldsymbol{m},\boldsymbol{m},\boldsymbol{k}} \widehat{u}^{(2)}_{\boldsymbol{n}} + a^{(j,3)}_{\boldsymbol{m},\boldsymbol{m},\boldsymbol{k}} \widehat{u}^{(3)}_{\boldsymbol{n}} = \lambda_N \sum_{\boldsymbol{n} \in \mathbb{I}_N} \widehat{\varepsilon}_{\boldsymbol{m}-\boldsymbol{n}} \widehat{\boldsymbol{u}}^{(j)}_{\boldsymbol{n}}$$

for $j = 1, 2, 3$ and all $\boldsymbol{m} \in \mathbb{I}_N$. Now we choose the ordering of the components of the Fourier coefficient vector the same way as for the H-field problem. This means we choose the ordering for the unknown coefficients by starting with all the first components, then all the second components and finally all the third components. This ordering leads to the generalized matrix eigenvalue problem

$$A\widehat{\boldsymbol{u}} = \lambda T \widehat{\boldsymbol{u}}, \tag{7.10}$$

where the symmetric matrix A consists of nine diagonal blocks (each of dimension $d := (2N+1)^3$), i.e.

$$A = \left(\begin{array}{c|c|c} \ddots & \ddots & \ddots \\ \hline \ddots & \ddots & \ddots \\ \hline \ddots & \ddots & \ddots \end{array} \right)$$

and T is a Hermitian block-diagonal matrix, which contains three times the block $[\![\varepsilon]\!]$ on the diagonal, i.e.

$$T = \begin{pmatrix} [\![\varepsilon]\!] & & \\ & [\![\varepsilon]\!] & \\ & & [\![\varepsilon]\!] \end{pmatrix}.$$

Notice that we now deal with a Toeplitz matrix $[\![\varepsilon]\!]$ which is generated by a function $\varepsilon : \mathbb{R}^3 \to \mathbb{R}$. This means that $[\![\varepsilon]\!]$ is not as in the two-dimensional case just a BTTB matrix, it is rather a BTTB matrix whose blocks are themselves BTTB matrices. This means we have a deeper nesting level of the BTTB structure. Matrices which are generated by some n-variate function $f : \mathbb{R}^n \to \mathbb{R}$ have n levels in their block structure. Matrices of this type are called *multilevel* BTTB matrices. For a general notion of multilevel matrices we refer to Chapter 2 in [53]. Our discretization leads us to a generalized eigenvalue problem with matrices A and T. Notice that A is sparse, thus a matrix-vector product can be performed in $\mathcal{O}(d)$ operations. The matrix T is not sparse, however matrix-vector products with T can be realized via FFT in $\mathcal{O}(d \log(d))$ operations due to its Toeplitz structure. For an introduction to iterative Toeplitz solvers we refer to [12]. Recent developments on the parallel solution of multilevel Toeplitz systems can be found in [13] and references therein.

7.3 Divergence constraint

So far we have only considered how the two curl problems (7.1) and(7.8) can be discretized with the Fourier-Galerkin method. We have not discussed how the divergence constraints

$$(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{h} = 0$$

and

$$(\nabla + \mathrm{i}\boldsymbol{k}) \cdot (\varepsilon \boldsymbol{e}) = 0$$

for the 3D problems can be realized in our discretizations. This is also discussed in [35]. Since any $\boldsymbol{u}_N \in \mathcal{T}_N$ can be represented as

$$\boldsymbol{u}_N(\boldsymbol{x}) = \sum_{\boldsymbol{n} \in \mathbb{I}_N} e^{\mathrm{i}2\pi \boldsymbol{n} \cdot \boldsymbol{x}} \widehat{\boldsymbol{u}}_{\boldsymbol{n}},$$

the divergence constraint

$$(\nabla + \mathrm{i}\boldsymbol{k}) \cdot \boldsymbol{u}_N = 0$$

translates into

$$(2\pi \boldsymbol{n} + \boldsymbol{k}) \cdot \widehat{\boldsymbol{u}}_{\boldsymbol{n}} = 0$$

for all $\boldsymbol{n} \in \mathbb{I}_N$. In practice this means that each of the Fourier coefficients only generates two degrees of freedom, because each of the Fourier coefficients $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$ is restricted to be an element of a two-dimensional subspace of $\mathbb{C}^3$ orthogonal to the space spanned by $2\pi \boldsymbol{n} + \boldsymbol{k}$. In practice this means we represent $\widehat{\boldsymbol{u}}_{\boldsymbol{n}}$ as a linear combination

$$\widehat{\boldsymbol{u}}_{\boldsymbol{n}} = \widetilde{u}^1_{\boldsymbol{n}} \boldsymbol{p}^1_{\boldsymbol{n}} + \widetilde{u}^2_{\boldsymbol{n}} \boldsymbol{p}^2_{\boldsymbol{n}}, \tag{7.11}$$

where $\boldsymbol{p}^1_{\boldsymbol{n}}$ and $\boldsymbol{p}^2_{\boldsymbol{n}}$ are two orthonormal basis vectors of the two-dimensional subspace of $\mathbb{C}^3$. Let us first consider the discretization of the H-field formulation, which resulted in the Hermitian eigenvalue problem (7.7), i.e.

$$B^\top A^\top T\, A\, B \widehat{\boldsymbol{u}} = \lambda \widehat{\boldsymbol{u}}.$$

We define $d := (2N+1)^3$. Then the dimension of the matrix of our eigenvalue problem is $3d \times 3d$. Next we define $P \in \mathbb{C}^{3d \times 2d}$ as the block-diagonal matrix, whose diagonal blocks contain the two basis vectors that result from (7.11). With P we can realize the divergence constraint. The eigenvalue problem (7.7) becomes

$$P^\top B^\top A^\top T\, A\, B P \widetilde{\boldsymbol{u}} = \lambda P^\top P \widetilde{\boldsymbol{u}} = \lambda \widetilde{\boldsymbol{u}}. \tag{7.12}$$

Notice that the eigenvalue problem previously was of dimension $3d \times 3d$ and now is of dimension $2d \times 2d$. Next we want to consider the discretized E-field problem (7.10), which reads

$$A \widehat{\boldsymbol{u}} = \lambda T \widehat{\boldsymbol{u}}.$$

For this problem the constraint was

$$(\nabla + \mathrm{i}\boldsymbol{k}) \cdot (\varepsilon \boldsymbol{e}) = 0.$$

This means that we have to apply P to the vector which contains the Fourier coefficients of $\varepsilon \boldsymbol{e}$. But in the discretization of the E-field problem the BTTB matrix T was generated by Fourier coefficients of ε and the vector $\widehat{\boldsymbol{u}}$ contained the Fourier coefficients of our approximate solution to $\boldsymbol{e}$. So the matrix-vector product $T\widehat{\boldsymbol{u}}$ is the convolution of the Fourier coefficients of ε with those of $\boldsymbol{e}$. This means that the vector $T\widehat{\boldsymbol{u}}$ can be interpreted as the vector which contains the Fourier coefficients of $\varepsilon \boldsymbol{e}$. This means

$$P\widetilde{\boldsymbol{u}} = T\widehat{\boldsymbol{u}}.$$

Moreover, the diagonal blocks $[\![\varepsilon]\!]$ of the matrix T are generated by the Fourier coefficients of a real valued, piecewise constant function which typically attains values between 1 and 14, due to the considered materials like air, glass or silicon. Since the generating function is real-valued, such a matrix is Hermitian. Tilli has shown in [68], that if a function $f : \mathbb{R}^n \to \mathbb{R}$ is in $L^1(\Omega)$, then the eigenvalues of a BTTB matrix $[\![f]\!]$, which is generated by the Fourier coefficients of f, can be bounded from below by $\operatorname{ess\,inf}_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$ and from above by $\operatorname{ess\,sup}_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$. In [68] this was proved for the ordinary BTTB case, however, as mentioned in the paper the same proof is applicable for the multilevel BTTB case. This result also follows from Theorem 2.2 in [61], where a more general case for singular values was proved. Since in our application $\varepsilon : \mathbb{R}^3 \to \mathbb{R}_{>0}$ is a two-valued function, we know that the minimal eigenvalue of $[\![\varepsilon]\!]$ is bounded from below by $\varepsilon_{\min}$ and the maximal eigenvalue is bounded from above by $\varepsilon_{\max}$. Consequently, the same is also true for T. This is the reason why the condition number of the matrix T is uniformly bounded by $\varepsilon_{\max}/\varepsilon_{\min}$, for any discretization level $N \in \mathbb{N}$ that was chosen in $\mathbb{I}_N$.

Theorem 7.3.1. *The condition number of the Hermitian matrix T in the discretization (7.10) stays uniformly bounded for all discretization levels $N \in \mathbb{N}$, and it holds*

$$cond_2(T) \leq \frac{\varepsilon_{\max}}{\varepsilon_{\min}}.$$

So T is always invertible and is always well-conditioned. This makes it well-suited for applying the cg-method for computing $T^{-1}\boldsymbol{x}$ for a given $\boldsymbol{x}$, since matrix-vector products can be executed efficiently via FFT. Now we can write the discretized E-field problem (7.10) as

$$AT^{-1}T\widehat{\boldsymbol{u}} = \lambda T\widehat{\boldsymbol{u}}.$$

Applying P finally leads to the non-symmetric eigenvalue problem

$$P^\top AT^{-1}P\widetilde{\boldsymbol{u}} = \lambda P^\top P\widetilde{\boldsymbol{u}} = \lambda\widetilde{\boldsymbol{u}}. \tag{7.13}$$

Notice that we now have a standard matrix eigenvalue problem and that the dimension of the problem again was reduced from $3d \times 3d$ to $2d \times 2d$. This way we also avoid to compute d zero eigenvalues, which we are not interested in. When we apply iterative methods we need to compute $P^\top A T^{-1} P \boldsymbol{x}$, for a given vector $\boldsymbol{x}$. This means that we have to solve linear systems with the coefficient matrix T. As discussed above, this can be done efficiently with the cg-method. However, it would be even more efficient to apply some representation for the inverse of T. For usual Toeplitz matrices such a concept was presented in [30]. Another step to find such representations for BTTB matrices recently was presented in [43]. In this paper about the inverses of BTTB matrices it was roughly stated that the inverses of BTTB matrices can be represented as a sum of products of two circulant matrices. This result was presented for the usual BTTB matrix case, which corresponds for our problems to the two-dimensional case. If we were able to do that for a three-level Block Toeplitz matrix $[\![\varepsilon]\!]$, this would be a huge improvement. On the one hand there would not be any need to embed the matrix $[\![\varepsilon]\!]$ into a circulant one of roughly twice the dimension, because we want to work with the FFT. If it is a circulant matrix, then we can work with the FFT directly, which makes products in 3D faster by a factor of roughly 8. Moreover, with a formula for an inverse we would compute directly $T^{-1}\boldsymbol{x}$ for a given $\boldsymbol{x}$, without iterations. Since T is well-conditioned the number of iterations with the *cg*-method is moderate. However, we need to multiply the number of iterations with the factor about which the matrix vector product with T is more time consuming than for a circulant matrix of the size of T. This means, that using special techniques for inverses of Toeplitz matrices would improve the algorithm. Applying such techniques in order to improve the algorithm is still ongoing work. It would be desirable to extend the result for BTTB matrices in [43] to three-level Block Toeplitz matrices.

7.4 Remark to inverse rule

In Chapter 5 we have discussed the so-called inverse rule. At first sight it does not seem to be desirable to apply this inverse convolution of Fourier coefficients. Let us discuss why this would be desirable in the discretization of the E-field formulation. If the inverse rule was applicable in any case, then with (7.10) we could consider

$$A\widehat{\boldsymbol{u}} = \lambda T^{-1}\widehat{\boldsymbol{u}},$$

with $T = \left[\!\left[\frac{1}{\varepsilon}\right]\!\right]$. Then one could apply $P\widetilde{\boldsymbol{u}} = T^{-1}\widehat{\boldsymbol{u}}$, with P as in Section 7.3. With

$$ATT^{-1}\widehat{\boldsymbol{u}} = \lambda T^{-1}\widehat{\boldsymbol{u}},$$

this would lead to

$$P^\top ATP\widetilde{\boldsymbol{u}} = \lambda P^\top P\widetilde{\boldsymbol{u}} = \lambda\widetilde{\boldsymbol{u}}.$$

This means, that the inverse T^{-1} would disappear in our iterations and we actually would not need to work with it. However, we have seen in the example in Section 5.5.1 that the application of the inverse rule should not be applied when the situations in the factorization theorems do not occur. As mentioned in [39], due to physical properties of the fields the product $(\varepsilon E_1)(x_1, x_2, x_3)$ is continuous in the variable x_1, which leads to the application of the inverse rule. In the other coordinates the Laurent rule needs to applied, because the products are discontinuous. However, the application of different rules in different directions leads in their approach to the situation that inverses of matrices have to be computed explicitly, which is very costly. In [39] it was not investigated how optimized implementations of their method and the usual one compare.

7.5 Convergence of the discretization

Due to the fact that discretizing the H-field formulation with Fourier methods leads to convergence problems, in [35] it was suggested to replace the permittivity function ε by a smoothed one, in order to obtain better convergence rates. However, in [52] the idea of smoothed potentials was discussed and analyzed, and [35] was explicitly mentioned in that discussion. It was shown that, due to the additional error that is being introduced by the smoothing, the overall convergence with Fourier methods is not better than without smoothing. Their conclusion in [52] was that smoothing of potentials is not worth it. In Section 8.5 we will consider a 3D example that was also considered in [10]. We will use eigenvalues from [10], which were computed on a parallel computer, as reference eigenvalues in order to compare the results of the H-field and E-field discretizations. The numerical results in Section 8.5 suggest that the E-field formulation is the better choice for discretizations with Fourier methods, if one can approximate solutions numerically in an efficient way. This was also observed in [64]. As already mentioned, in [39] better convergence results were obtained, at the cost of computing explicitly inverses. Recent results in [18] show that the solution of the H-field formulation has a piecewise higher regularity than that of the E-field formulation. This can be exploited in finite element approximations, in contrast to approximations with the Fourier-Galerkin method, because it depends on the global regularity. This legitimates the question, whether finite elements are the better choice for 3D photonic band structure computations, since it has more local flexibility due to grid refinement and the choice of local polynomials.

Chapter 8

Eigenvalue Solver

In this final chapter we discuss the solution of the eigenvalue problems that arise from the 3D problems that were discretized in Chapter 7. For a thorough and general treatment of numerical methods for matrix eigenvalue problems we refer to the standard textbook by Watkins [71], where most of the concepts introduced in this chapter can be found. Other standard references for the numerical treatment of eigenvalue problems are [6, 58].

8.1 Matrix eigenvalue problem

In this chapter we will deal with the numerical solution of matrix eigenvalue problems with large dimension. So let $C \in \mathbb{C}^{n\times n}$ with $n \gg 1$. We want to find $\boldsymbol{u} \in \mathbb{C}^n \backslash \{\mathbf{0}\}$ and $\lambda \in \mathbb{C}$ such that

$$C\boldsymbol{u} = \lambda \boldsymbol{u}.$$

Actually, with regard to the discretizations in Chapter 7 we are interested in computing a few smallest eigenvalues of a matrix of large dimension. We will now introduce the concepts for a more general case and postpone the discussion of how to compute a few smallest eigenvalues to Section 8.4.

8.2 Ritz and harmonic Ritz values

In the next two sections we will explain the basics that we need for the numerical approximation of eigenvalues of matrices. We will stick to the very nice presentation of the topic in [32], where all the concepts presented here are discussed in much more detail and where all the proofs of the following theorems can be found. Another detailed discussion of the topics that follow can be found in Chapter 9 of

[71]. We will skip the proofs, because we only want to collect the most important results for our following discussion. We start with a definition for approximate eigenvalues and eigenvectors of a matrix.

Definition 8.2.1 ([32], Definition 2.7)**.** *Given a subspace $\mathcal{S} \subseteq \mathbb{C}^n$, $\theta_k \in \mathbb{C}$ is a **Ritz value** of C with respect to $\mathcal{S}$ with **Ritz vector** $\boldsymbol{u}_k$ if $(\theta_k, \boldsymbol{u}_k)$ satisfies the **Galerkin condition***

$$\boldsymbol{u}_k \in \mathcal{S}, \quad \boldsymbol{u}_k \neq \boldsymbol{0}, \qquad C\boldsymbol{u}_k - \theta_k \boldsymbol{u}_k \perp \mathcal{S}. \tag{8.1}$$

*The pair $(\theta_k, \boldsymbol{u}_k)$ is called a **Rayleigh-Ritz approximation**.*

For practical reasons it would be nice to have a characterization of a Rayleigh-Ritz approximation which is more concrete. The following theorem provides such a characterization via certain matrices.

Theorem 8.2.2 ([32], Theorem 2.8)**.** *Let $\mathcal{S} \subseteq \mathbb{C}^n$ be an m-dimensional subspace and $V \in \mathbb{C}^{n\times m}$ be any matrix such that $\mathcal{R}(V) = \mathcal{S}$. Then $(\theta_k, \boldsymbol{u}_k)$ is a Rayleigh-Ritz approximation if and only if θ_k is an eigenvalue of*

$$H = (V^{\mathrm{H}}V)^{-1}V^{\mathrm{H}}CV$$

and $\boldsymbol{u}_k = V\boldsymbol{y}_k$, where $\boldsymbol{y}_k$ is an eigenvector of H corresponding to θ_k.

Notice that the dimension of the matrix H is $m \times m$, where $m \in \mathbb{N}$ is much smaller than $n \in \mathbb{N}$. This means that the calculation of the Ritz values of H is feasible with standard algorithms. Moreover, considering H we realize that choosing an orthonormal basis of $\mathcal{S}$ the matrix H reduces to $H = V^{\mathrm{H}}CV$.
With Algorithm 8.1 one can compute Rayleigh-Ritz approximations. By combining the Rayleigh-Ritz method with an algorithm that constructs a nested sequence of subspaces

$$\mathcal{S}_{m+1} = \mathcal{S}_m + \operatorname{span}\{\boldsymbol{v}_{m+1}\}, \qquad \mathcal{S}_1 = \operatorname{span}\{\boldsymbol{v}_1\},$$

with a suitably chosen sequence of vectors $\{\boldsymbol{v}_m\}_m$, one can construct efficient algorithms. Depending on the choice of the vectors $\boldsymbol{v}_m$ and the construction of a suitable basis of $\mathcal{S}_m$ the algorithms distinguish one from the other. It makes sense to utilize the property that the subspaces are nested. This means if the columns of $V \in \mathbb{R}^{n\times m}$ are a basis of $\mathcal{S}_m$, then we set

$$V_{m+1} = [V_m\, \boldsymbol{v}_{m+1}], \quad \boldsymbol{v}_{m+1} \in \mathcal{S}_{m+1}\backslash\mathcal{S}_m.$$

Numerically it makes sense to choose $\boldsymbol{v}_{m+1}$ in a way such that V_{m+1} is unitary, i.e. the basis is orthonormal. For Hermitian matrices one can prove the following result with the Min-Max theorem.

Algorithm 8.1 Rayleigh-Ritz method

Given C and a basis S of $\mathcal{S}$.

1. Orthonormalize the columns of S to obtain a unitary $V \in \mathbb{C}^{n\times m}$ with $\mathcal{R}(V) = \mathcal{S}$.
2. Form CV by m calls to a function providing a matrix-vector multiplication $\boldsymbol{x} \mapsto C\boldsymbol{x}$.
3. Form the Rayleigh-quotient matrix $H = V^{\mathrm{H}}CV$.
4. Compute the $k \leq m$ wanted eigenpairs $(\theta_j, \boldsymbol{y}_j)$ of H.
5. If necessary, compute the Ritz vectors $\boldsymbol{u}_j = V\boldsymbol{y}_j$.
6. Compute k residuals $\boldsymbol{r}_j = C\boldsymbol{u}_j - \theta_j\boldsymbol{u}_j = (CV)\boldsymbol{y}_j - \theta_j\boldsymbol{u}_j$.

Theorem 8.2.3 ([32], Theorem 2.16)**.** *Assume that $C = C^{\mathrm{H}}$ and let $\{\mathcal{S}_m\}_m$ be a nested sequence $(\mathcal{S}_m \subseteq \mathcal{S}_{m+1})$ of subspaces with $\dim \mathcal{S}_m = m$. Denote by $\theta_k^{(m)}$, $k = 1, ..., m$, the Ritz values corresponding to $\mathcal{S}_m$ and by λ_k, $k = 1, ..., n$, the eigenvalues of C, all ordered in decreasing order, i.e. $\theta_m^{(m)} \leq \cdots \leq \theta_1^{(m)}$ for all m and $\lambda_n \leq \cdots \leq \lambda_1$. Then*

$$\theta_k^{(m)} \leq \theta_k^{(m+1)} \leq \lambda_k \qquad \text{and} \qquad \lambda_{n-k+1} \leq \theta_{m+1-k+1}^{(m+1)} \leq \theta_{m-k+1}^{(m)}.$$

This theorem can be interpreted in a way such that *outer* eigenvalues can be approximated well, while *inner* eigenvalues are not so easy to approximate. If one is interested in inner eigenvalues then one can use a shift σ and work with $(C-\sigma)^{-1}$ instead of C. However, this makes it necessary to solve linear systems. In order to avoid costly solutions of linear systems the concept of *harmonic Ritz values* was introduced. In the following we discuss this concept for $\sigma = 0$, because in our band structure calculations we are interested in a few lowest magnitude eigenvalues of a matrix of large dimension.

Definition 8.2.4 ([32], Definition 2.17)**.** *Given a subspace $\mathcal{S} \subseteq \mathbb{C}^n$, $\theta_k \neq 0$ is a* ***harmonic Ritz value*** *of C with respect to $\mathcal{S}$ if θ_k^{-1} is a Ritz value of C^{-1} with respect to $\mathcal{S}$.*

Now our problem has changed from computing a few lowest magnitude eigenvalues of a matrix C to computing a few largest magnitude eigenvalues of the matrix C^{-1}. The next theorem tells us how we can avoid working with C^{-1}.

Theorem 8.2.5 ([32], Theorem 2.18)**.** *Let $\mathcal{V} \subseteq \mathbb{C}^n$ be a subspace with $\dim \mathcal{V} = m$, and orthonormal basis $V \in \mathbb{C}^{n\times m}$. Assume that $\mathcal{W} = C\mathcal{V}$ has dimension m. Then*

θ_k is a harmonic Ritz value of C with respect to $\mathcal{W}$ if and only if

$$C\boldsymbol{u}_k - \theta_k \boldsymbol{u}_k \perp \mathcal{W} \qquad \text{for some} \qquad \boldsymbol{u}_k \in \mathcal{V}, \boldsymbol{u}_k \neq \boldsymbol{0}. \tag{8.2}$$

Moreover, if $W \in \mathbb{C}^{n\times m}$ is a matrix with $\mathcal{R}(W) = \mathcal{W}$, then (8.2) is equivalent to

$$(W^{\mathrm{H}}CV)\boldsymbol{y}_k = \theta_k (W^{\mathrm{H}}V)\boldsymbol{y}_k, \qquad \boldsymbol{u}_k = V\boldsymbol{y}_k. \tag{8.3}$$

Such a vector $\boldsymbol{u}_k$ is called a *harmonic Ritz vector* associated with the harmonic Ritz value θ_k and $(\theta_k, \boldsymbol{u}_k)$ a *harmonic Rayleigh-Ritz approximation*. If $W^{\mathrm{H}}V$ is nonsingular, then the generalized eigenvalue problem (8.2) can be reformulated into the equivalent ordinary eigenvalue problem

$$G\boldsymbol{y}_k = \theta_k \boldsymbol{y}_k, \qquad G = (W^{\mathrm{H}}V)^{-1}(W^{\mathrm{H}}CV), \qquad \boldsymbol{u}_k = V\boldsymbol{y}_k. \tag{8.4}$$

8.3 Arnoldi iteration

In this section we discuss the choice of nested subspaces $\{\mathcal{S}_m\}_m$, which have turned out to be a good choice for the resulting algorithms. We start with the following definition.

Definition 8.3.1. *For a given $C \in \mathbb{C}^{n\times n}$ and nonzero $\boldsymbol{b} \in \mathbb{C}^n$,*

$$\mathcal{K}_m(C, \boldsymbol{b}) = \operatorname{span}\{\boldsymbol{b}, C\boldsymbol{b}, ..., C^{m-1}\boldsymbol{b}\}$$

*is called the m-th **Krylov subspace** with respect to C and $\boldsymbol{b}$.*

We want to apply the Rayleigh-Ritz method to a sequence of Krylov subspaces. Choosing $\{C^j\boldsymbol{b}\}_{j=1}^{m-1}$ as a basis is not a good choice, because the vectors become numerically linearly dependent. This means we have to find a better way for constructing a basis of $\mathcal{K}_m = \mathcal{K}_m(C, \boldsymbol{b})$. Due to the relation $\mathcal{K}_m \subseteq \mathcal{K}_{m+1}$, we only need to compute one new basis vector at each iteration step.

Lemma 8.3.2 ([32], Lemma 2.22)**.** *If $V_j \in \mathbb{C}^{n\times j}$ is a basis of $\mathcal{K}_j$, $j = 1, ..., m+1$, where $V_j = [V_{j-1}\, \boldsymbol{v}_j]$, then there is a unique unreduced upper Hessenberg matrix $\widetilde{H}_m = (h_{ij}) \in \mathbb{C}^{m+1\times m}$ such that*

$$CV_m = V_{m+1}\widetilde{H}_m. \tag{8.5}$$

It can be shown that equation (8.5) is equivalent to

$$CV_m = V_m H_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^{\top}, \tag{8.6}$$

where

$$H_m = [I_m \ \mathbf{0}]\widetilde{H}_m \in \mathbb{C}^{m\times m} \qquad \text{or} \qquad \widetilde{H}_m = \left[\begin{array}{c} H_m \\ \hline \mathbf{0} \ h_{m+1,m} \end{array}\right].$$

In the case that $h_{m+1,m} = 0$ holds, the Krylov subspace $\mathcal{K}_m = \mathcal{R}(V_m)$ is an C-invariant subspace and $\lambda(H_m) \subseteq \lambda(C)$. In general, this only holds when $m = n$. In the following Arnoldi algorithm an orthonormal basis of $\mathcal{K}_m$ is being constructed:

$$\boldsymbol{v}_1 = \boldsymbol{b}/\left\|\boldsymbol{b}\right\|, \qquad \boldsymbol{v}_{m+1} = \frac{(I - V_m V_m^{\mathrm{H}})C\boldsymbol{v}_m}{\|(I - V_m V_m^{\mathrm{H}})C\boldsymbol{v}_m\|}, \qquad m = 1, 2, ...$$

In the m-th step the vector $C\boldsymbol{v}_m$ is being orthogonalized against all the previous basis vectors with the *modified Gram-Schmidt* process. A numerically stable version is given in the following algorithm.

Algorithm 8.2 Arnoldi algorithm with modified Gram-Schmidt

Given $C \in \mathbb{C}^{n\times n}$, $\boldsymbol{b} \in \mathbb{C}^n$, and $\beta = \|\boldsymbol{b}\| > 0$.

$\boldsymbol{v}_1 = \boldsymbol{b}/\beta$
for $m = 1, 2, \ldots$
(1) $\widetilde{\boldsymbol{v}}_{m+1} = C\boldsymbol{v}_m$
(2) for $j = 1, ..., m$
$h_{j,m} = \boldsymbol{v}_j^{\mathrm{H}}\widetilde{\boldsymbol{v}}_{m+1}$
$\widetilde{\boldsymbol{v}}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1} - h_{j,m}\boldsymbol{v}_j$
(3) $h_{m+1,m} = \|\widetilde{\boldsymbol{v}}_{m+1}\|$
(4) $\boldsymbol{v}_{m+1} = \widetilde{\boldsymbol{v}}_{m+1}/h_{m+1,m}$

The next lemma gives some insight to the properties of the matrices that arise in the Arnoldi algorithm.

Theorem 8.3.3 ([32], Lemma 2.23)**.** *Let V_m and $\widetilde{H}_m$ be the matrices generated by the Arnoldi process. Then*

a) $CV_m = V_{m+1}\widetilde{H}_m = V_m H_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^{\top}$,

b) $V_m^{\mathrm{H}} C V_m = H_m$,

c) $V_{m+1}^{\mathrm{H}} C V_m = \widetilde{H}_m$,

d) If $C = C^{\mathrm{H}}$, then $H_m = H_m^{\mathrm{H}}$ is tridiagonal.

Now we want to combine the Arnoldi process with the Rayleigh-Ritz method. Theorem 8.2.2 tells us that the eigenvalues of H_m actually are the Ritz values, and that the Ritz vectors can be constructed with the eigenvectors of H_m. The next theorem tells us that we can compute the residual norms of Rayleigh-Ritz

approximations without additional costs. Moreover, the following lemma tells us that all residuals of m Rayleigh-Ritz approximations are scalar multiples of the next Arnoldi vector $\boldsymbol{v}_{m+1}$.

Lemma 8.3.4 ([32], Lemma 2.24)**.** *Let $(\theta_k, \boldsymbol{u}_k)$, $\boldsymbol{u}_k = V_m \boldsymbol{y}_k$ be Rayleigh-Ritz approximations of C corresponding to $\mathcal{K}_m$. Then the residual is given by*

$$C\boldsymbol{u}_k - \theta_k \boldsymbol{u}_k = h_{m+1,m}(\boldsymbol{e}_m^\top \boldsymbol{y}_k)\boldsymbol{v}_{m+1}$$

and thus $\|C\boldsymbol{u}_k - \theta_k \boldsymbol{u}_k\| = h_{m+1,m}|\boldsymbol{e}_m^\top \boldsymbol{y}_k|$.

In Definition 8.2.4 and Theorem 8.2.5 we have introduced the concept of harmonic Ritz values. The reason is that we want to compute harmonic Ritz values for the problems discussed in Section 7.3. Therefore, we discuss in the following how the Arnoldi process can be applied for computing harmonic Ritz values. By Theorem 8.2.5, we have to solve the generalized eigenvalue problem

$$(W_m^{\mathrm{H}} C V_m)\boldsymbol{y}_k = \theta_k (W_m^{\mathrm{H}} V_m)\boldsymbol{y}_k.$$

Since the size of the problem is $m \ll n$, it can be solved easily with standard algorithms. If V_m, H_m are the matrices from the Arnoldi process, then with $W_m = CV_m$ and Theorem 8.3.3 we obtain

$$W_m^{\mathrm{H}} V_m = V_m^{\mathrm{H}} C^{\mathrm{H}} V_m = H_m^{\mathrm{H}}.$$

With the Arnoldi recursion (8.6) it follows

$$W^{\mathrm{H}} C V_m = (CV_m)^{\mathrm{H}} C V_m = (V_{m+1}\widetilde{H}_m)^{\mathrm{H}}(V_{m+1}\widetilde{H}_m) = \widetilde{H}_m^{\mathrm{H}}\widetilde{H}_m.$$

Finally, we end up with

$$(\widetilde{H}_m^{\mathrm{H}}\widetilde{H}_m)\boldsymbol{y}_k = \theta_k H_m^{\mathrm{H}} \boldsymbol{y}_k. \tag{8.7}$$

For the condition number it holds $\kappa(\widetilde{H}_m^{\mathrm{H}}\widetilde{H}_m) = \kappa(\widetilde{H}_m)^2$. Therefore, it is advisable to avoid working with $\widetilde{H}_m^{\mathrm{H}}\widetilde{H}_m$. With the identity

$$\widetilde{H}_m = \left[\begin{array}{c} H_m \\ \hline \mathbf{0}\ h_{m+1,m} \end{array}\right]$$

we obtain

$$\widetilde{H}_m^{\mathrm{H}}\widetilde{H}_m = H_m^{\mathrm{H}} H_m + h_{m+1,m}^2 \boldsymbol{e}_m \boldsymbol{e}_m^\top.$$

Therefore, equation (8.7) becomes

$$\left(H_m^{\mathrm{H}} H_m + h_{m+1,m}^2 \boldsymbol{e}_m \boldsymbol{e}_m^\top\right)\boldsymbol{y}_k = \theta_k H_m^{\mathrm{H}} \boldsymbol{y}_k.$$

Assuming that H_m is nonsingular, with

$$\boldsymbol{f} := H_m^{-\mathrm{H}} \boldsymbol{e}_m, \tag{8.8}$$

we obtain the eigenvalue problem

$$\left(H_m + h_{m+1,m}^2 \boldsymbol{f}\boldsymbol{e}_m^\top\right)\boldsymbol{y}_k = \theta_k \boldsymbol{y}_k. \tag{8.9}$$

8.4 Harmonic Restarted Arnoldi

Since we are interested in computing the lowest magnitude eigenvalues of a matrix, without working with the inverse of that matrix, we want to compute its harmonic Ritz values. A serious issue of the Arnoldi process one has to deal with is the fact that expense and storage grow with the number of iterations, i.e. as the Krylov subspace increases. This can lead to the undesired situation that the eigenvalues we are interested in cannot be approximated well enough, due to limitations on our computational resources. This is the reason why a so-called *restarting* of the Arnoldi recurrence becomes necessary in practice. In order to compute several eigenvalues simultaneously, it is necessary to retain several approximate eigenvectors. The so-called *implicitly restarted Arnoldi method (IRA)* proposed by Sørensen [63] solves this problem and is a standard tool nowadays, which is also built into the `eigs` package in MATLAB®. Here we only want to mention the most important facts of IRA that we need. This will serve as a preparation for a numerical method for computing harmonic Ritz values which we are interested in for our problem, namely the method presented in [49]. For a detailed discussion of IRA we refer to the original work [63]. Morgan showed in Theorem 3 of [47] that the subspace generated by IRA is

$$\text{span}\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k, \boldsymbol{v}_{m+1}, C\boldsymbol{v}_{m+1}, C^2\boldsymbol{v}_{m+1}, C^3\boldsymbol{v}_{m+1}, ..., C^{m-k-1}\boldsymbol{v}_{m+1}\}, \tag{8.10}$$

where $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k$ are Ritz vectors from the previous cycle and $\boldsymbol{v}_{m+1}$ is the $(m+1)$st Arnoldi vector from the previous cycle. Moreover, in [47] it was shown that the IRA subspace (8.10) is the same space as

$$\text{span}\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k, C\boldsymbol{y}_i, C^2\boldsymbol{y}_i, C^3\boldsymbol{y}_i, ..., C^{m-k}\boldsymbol{y}_i\}, \tag{8.11}$$

for each $1 \leq i \leq k$. This illustrates why IRA works so well. As mentioned in [47], the IRA subspace contains a Krylov subspace of dimension $m-k+1$ with each of the desired Ritz vectors as a starting vector. This allows us to approximate all the desired eigenpairs simultaneously with IRA. In our numerical example for a 3D photonic crystal we will use the so-called *Harmonic Restarted Arnoldi (HRA)* method by Morgan and Zeng [49]. This method was developed to realize a restart technique while computing harmonic Ritz values. Here we want to summarize the most important facts about the HRA method that we need for our numerical band structure computations in 3D. For a detailed discussion of the HRA method we refer to the original work [49]. Here we will only reiterate the essentials that we need. Actually, in [49] it is discussed for a more general setting with shifts which lie in a region where eigenvalues are desired. As we here are interested in the lowest magnitude eigenvalues, this corresponds to a shift which is zero. We will discuss only this case. The subspace generated by HRA is

$$\text{span}\{\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, ..., \widetilde{\boldsymbol{y}}_k, \boldsymbol{r}, C\boldsymbol{r}, C^2\boldsymbol{r}, C^3\boldsymbol{r}, ..., C^{m-k-1}\boldsymbol{r}\}, \tag{8.12}$$

where $\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, ..., \widetilde{\boldsymbol{y}}_k$ are harmonic Ritz vectors from the previous cycle and $\boldsymbol{r}$ is a multiple of the harmonic residual vectors [49]. In [46] for the residuals it was shown that

$$C\widetilde{\boldsymbol{y}}_i - \widetilde{\theta}_i\widetilde{\boldsymbol{y}}_i = \gamma_i \boldsymbol{r} \tag{8.13}$$

holds, for some scalars γ_i. This means that the residuals are multiples one of each other. Moreover, in [46] it was shown that the whole subspace (8.12) is a Krylov subspace and that it is the same space as

$$\text{span}\{\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2, ..., \widetilde{\boldsymbol{y}}_k, C\widetilde{\boldsymbol{y}}_i, C^2\widetilde{\boldsymbol{y}}_i, C^3\widetilde{\boldsymbol{y}}_i, ..., C^{m-k}\widetilde{\boldsymbol{y}}_i\}, \tag{8.14}$$

for each $1 \leq i \leq k$. This means that the subspace has similar properties to those of IRA discussed above. It has the property that it contains a Krylov subspace of dimension $m - k + 1$ with each of the desired harmonic Ritz vectors as a starting vector. The HRA algorithm by Morgan and Zeng [49] is given in Algorithm 8.3. In the following discussion we will give some explanations from [49] to the algorithm. The HRA algorithm, similarly to the previous discsussion, generates a recurrence of the form

$$CV_m = V_{m+1}\widetilde{H}_m = V_m H_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^\top. \tag{8.15}$$

In step 1 of Algorithm 8.3 the maximum size of the subspace, the number of approximate eigenvectors to be retained from the previous cycle, the desired number of eigenpairs and an initial vector for the iteration is chosen. In step 2 the Arnoldi iteration is being applied. In step 3 the the eigenvalue problem of small dimension is being solved. In step 4 of Algorithm 8.3 the residual norms need to be computed. In [48] it was shown that if $(\rho_i, \widetilde{\boldsymbol{y}}_i)$ is an approximate eigenpair, then the residual norm can be computed via

$$\|C\widetilde{\boldsymbol{y}}_i - \rho_i\widetilde{\boldsymbol{y}}_i\| = \sqrt{\|(H_m - \rho_i I)\widetilde{\boldsymbol{g}}_i\|^2 + h_{m+1,m}^2(\boldsymbol{e}_m^\top\widetilde{\boldsymbol{g}}_i)^2}. \tag{8.16}$$

If the computed residual norm is not smaller than a prescribed tolerance, then the restart is being prepared. Therefore, in step 5 one orthonormalizes the short vectors computed in step 3. Those are the approximate eigenvectors which we retain in the restarted iteration. Next, we have to choose $\boldsymbol{r}$ from (8.12) and (8.13), in order to run the new Arnoldi iteration. This is what happens in step 6 of Algorithm 8.3, and we want to discuss this now. With the recurrence from (8.15) and the identity (8.9) we obtain

$$\begin{aligned} C\widetilde{\boldsymbol{y}}_i - \widetilde{\theta}_i\widetilde{\boldsymbol{y}}_i &= CV_m\widetilde{\boldsymbol{g}}_i - \widetilde{\theta}_i V_m\widetilde{\boldsymbol{g}}_i \\ &\overset{(8.15)}{=} (V_m H_m + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^\top)\widetilde{\boldsymbol{g}}_i - \widetilde{\theta}_i V_m\widetilde{\boldsymbol{g}}_i \\ &= V_m(H_m - \widetilde{\theta}_i I)\widetilde{\boldsymbol{g}}_i + h_{m+1,m}\boldsymbol{v}_{m+1}\boldsymbol{e}_m^\top\widetilde{\boldsymbol{g}}_i \end{aligned}$$

Algorithm 8.3 Harmonic Restarted Arnoldi (HRA) [49]

1. *Start*: Choose m, the maximum size of the subspace, and k, the number of approximate eigenvectors that are retained from one cycle to the next. Also pick *numev*, the desired number of eigenpairs. Choose an initial vector $\boldsymbol{v}_1$ of unit length.

2. *Arnoldi iteration*: Apply the Arnoldi iteration from the current point to form the rest of V_{m+1} and $\widetilde{H}_m$. The current point is either from $\boldsymbol{v}_1$ if it is the first cycle or from $\boldsymbol{v}_{k+1}$ on the other cycles.

3. *Small eigenvalue problem*: Compute eigenpairs $(\widetilde{\theta}_i, \widetilde{\boldsymbol{g}}_i)$, with $\widetilde{\boldsymbol{g}}_i$ normalized, of $\left(H_m + h_{m+1,m}^2 \boldsymbol{f}\boldsymbol{e}_m^\top\right)\widetilde{\boldsymbol{g}} = \theta\widetilde{\boldsymbol{g}}$, where $\boldsymbol{f} = H^{-\mathrm{H}}\boldsymbol{e}_m$. Order the eigenpairs so that the first k are the desired ones. They normally would be the ones with $\widetilde{\theta}_i$'s nearest to 0. If desired, the harmonic Rayleigh quotients can be computed: $\rho_i = \widetilde{\boldsymbol{g}}_i^{\mathrm{H}} H_m \widetilde{\boldsymbol{g}}_i$.

4. *Check convergence*: Residual norms can be computed with (8.16) and convergence can be checked. If all desired eigenvalues have acceptable residual norm, then stop, first computing eigenvectors, if desired, as $\widetilde{\boldsymbol{y}}_i = V_m \widetilde{\boldsymbol{g}}_i$. Otherwise continue.

5. *Orthonormalization of first k short vectors*: In this step the restart begins. Orthonormalize $\widetilde{\boldsymbol{g}}_i$'s, for $1 \le i \le k$, in order to form a real m by k matrix P_k.

6. *Orthonormalization of the $k+1$ short vector*: Extend $p_1, ..., p_k$ to length $m+1$ by appending a zero to each, then orthonomalize $s = (-h_{m+1,m} f^\top, 1)^\top$ against $p_1, ..., p_k$ to form p_{k+1}. P_{k+1} is $m+1$ by $k+1$.

7. *Form portions of new H and V using the old H and V* : Let $\widetilde{H}_k^{new} = P_{k+1}^\top \widetilde{H}_m P_k$ and $V_{k+1}^{new} = V_{m+1} P_{k+1}$. Then let $\widetilde{H}_k = \widetilde{H}_k^{new}$ and $V_{k+1} = V_{k+1}^{new}$.

8. *Reorthogonalization of long k + 1 vector*: Orthogonalize $\boldsymbol{v}_{k+1}$ against the earlier columns of the new V_{k+1}. Go to step 2.

$$\begin{aligned} &\overset{(8.9)}{=} && -h_{m+1,m}^2 V_m \boldsymbol{f} \boldsymbol{e}_m^\top \widetilde{\boldsymbol{g}}_i + h_{m+1,m} \boldsymbol{v}_{m+1} \boldsymbol{e}_m^\top \widetilde{\boldsymbol{g}}_i \\ &= && h_{m+1,m} \boldsymbol{e}_m^\top \widetilde{\boldsymbol{g}}_i (\boldsymbol{v}_{m+1} - h_{m+1,m} V_m \boldsymbol{f}). \end{aligned}$$

Since we can write

$$\boldsymbol{v}_{m+1} - h_{m+1,m} V_m \boldsymbol{f} = V_{m+1} \begin{bmatrix} -h_{m+1,m} \boldsymbol{f} \\ 1 \end{bmatrix},$$

with

$$\boldsymbol{s} := \begin{bmatrix} -h_{m+1,m} \boldsymbol{f} \\ 1 \end{bmatrix}$$

we obtain

$$C \widetilde{\boldsymbol{y}}_i - \widetilde{\theta}_i \widetilde{\boldsymbol{y}}_i = h_{m+1,m} \boldsymbol{e}_m^\top \widetilde{\boldsymbol{g}}_i (V_{m+1} \boldsymbol{s}).$$

This means that one can choose $\boldsymbol{r}$ as $V_{m+1}\boldsymbol{s}$. With the vector $\boldsymbol{s}$ one forms the $(k+1)$st column of the new H in step 6 of Algorithm 8.3. In step 7 the new H and V are built. Finally, in step 8 one orthogonalizes the last column of the new V against the previous ones and starts with step 2 again. The total cost of operations is roughly $m^2 + km - k^2$ vector operations plus $(m-k)$ matrix-vector products per cycle. In the next section we will use this algorithm for a 3D band structure computation. We apply HRA with C chosen as the matrices on the left hand side in (7.12) and (7.13).

8.5 Numerical example for a 3D photonic crystal

In this section we want to consider a 3D example of a photonic crystal. It is the same example that was considered in [10, 39, 55, 64]. We consider the structure depicted in Figure 8.1. The width of the bars is 0.125 and the crystal function, which represents the considered crystal, attains the value 13 in the dark region, and the value 1 else. For the band structure computation we vary $\boldsymbol{k}$ along the path Γ–X–M–R in the Brillouin zone $B = [-\pi, \pi]^3$, as it is depicted in Figure 8.2. The computed band structure with the E-field formulation is depicted in Figure 8.3. Moreover, we want to compare results of the H-field formulation with those of the E-field formulation. For this we use the eigenvalues from page 100 in [10], which were computed on a parallel computer with finite elements, as reference values. In Table 8.1 and 8.2 we can see the corresponding errors of the first five eigenvalues to the reference eigenvalues, for the two different discretizations. If we compare the errors in Table 8.1 and 8.2 we see that for the same discretization order we

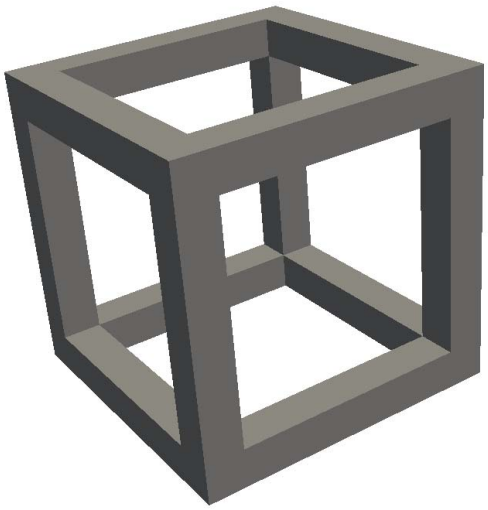


Figure 8.1: A 3D photonic crystal which attains the value 13 in the dark region, and the value 1 else. The width of the bars is 0.125.

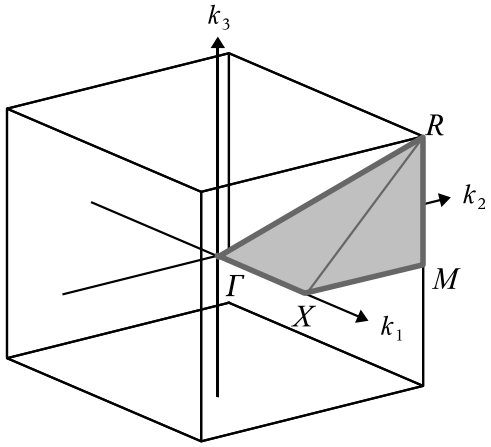


Figure 8.2: The path Γ–X–M–R–Γ in the Brillouin zone B.

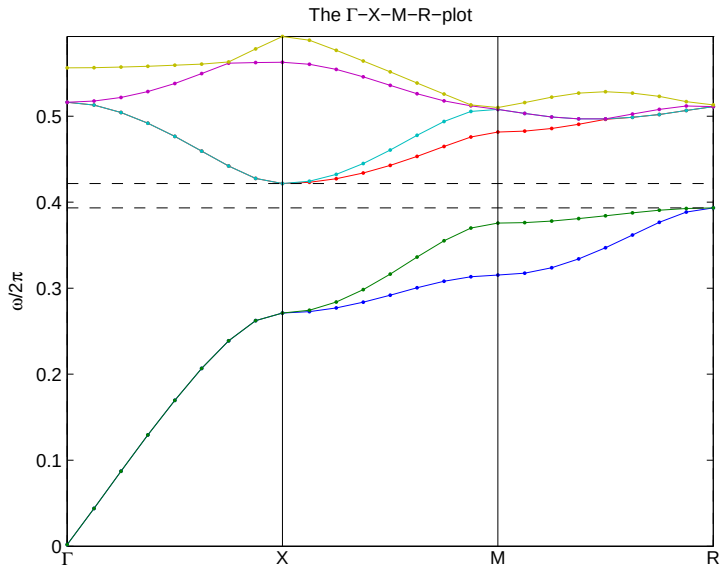


Figure 8.3: Band structure ($N = 15$) for a 3D photonic crystal.

obtain more accurate results with the E-field formulation. If we look at the first eigenvalue, we observe that there is a factor of more than 17 in the errors between the two discretizations. With those errors we have observed for both discretizations the experimental order of convergence to be one with respect to N. However, if we have a look at the errors for the E-field formulation and the errors for the H-field formulation, we observe that we obtain for the former more accurate results with $N = 31$ than with $N = 63$ for the latter. The same was observed in the physics literature [64]. In number of unknowns this reads 500.094 unknowns for the E-field formulation against 4.096.766 unknowns for the H-field formulation. These observations suggest, that although the experimental order of convergence is the same for both formulations, the E-field formulation is the better choice for discretizations with the Fourier-Galerkin method due the lower errors for the same discretization levels. This is interesting with regard to recent results for discretizations of the Maxwell equations with finite elements. In [18] it was shown that the solution of the H-field formulation has a piecewise higher regularity than that of the E-field formulation. The convergence of the finite element method only depends on the piecewise regularity. This is why in [18] it was concluded that discretizing the H-field formulation with finite elements is the better choice. In contrast to that, the Fourier-Galerkin method can not exploit the higher piecewise regularity, it depends on the global regularity of the solution. This can explain why in our case the discretization of the H-field formulation does not yield better results. In

contrast to finite element methods, our numerical results suggest that the E-field formulation is the better choice for the Fourier-Galerkin method, with respect to the number of degrees of freedom. In a test for the running time of the algorithms we computed 5 eigenvalues. We chose in all the tests the parameters in Algorithm 8.3 to be $m = 70$, $k = 20$ and the error tolerance as 10^{-4}. For the H-field formulation the problem, with $M = 15$ (59.582 unknowns), was solved after 25 Iterations in approximately 1 minute. With $M = 31$ (500.094 unknowns) it was solved after 65 Iterations in approximately 24 minutes. In the E-field formulation the problem, with $M = 15$ (59.582 unknowns), was solved after 23 Iterations in approximately 66 minutes. With $M = 31$ (500.094 unknowns) it was solved after 55 Iterations in approximately 20 hours. Clearly, the E-field formulation yields more accurate results, as it was already realized in [64]. But it is also clear that in our approach the E-field formulation is also much more costly in computational time. This is why we were interested in explicit representations of inverses of BTTB matrices as product of BCCB matrices in [43]. If such a representation could be extended to the three-level BTTB case, and it was usable it would make the E-field formulation more efficient, because no iterations would have to be performed for the computation of $T^{-1}\boldsymbol{x}$ for a given $\boldsymbol{x}$. In addition, the matrix-vector products via FFT would be faster since no embedding into a circulant matrix would be necessary. However, we were not yet able to make improvements into this direction. Moreover, there is space for improvement in the application of the HRA algorithm. We have realized in our computations, that the choice of the parameters can influence the performance.

Eigenvalue errors via E-field formulation					
Eigenvalue	reference	$N = 3$	$N = 7$	$N = 15$	$N = 31$
#1	3.9499	0.1185	0.0586	0.0284	0.0140
#2	4.7076	0.1216	0.0607	0.0300	0.0150
#3	8.5937	0.7767	0.3736	0.1791	0.0885
#4	9.5824	0.9313	0.4547	0.2172	0.1067
#5	10.9749	1.0627	0.5235	0.2488	0.1219

Table 8.1: Errors for the E-field formulation.

Eigenvalue errors via H-field formulation						
Eigenvalue	reference	$N = 3$	$N = 7$	$N = 15$	$N = 31$	$N = 63$
#1	3.9499	2.0651	0.7449	0.3218	0.1498	0.0722
#2	4.7076	2.3239	0.8519	0.3689	0.1718	0.0827
#3	8.5937	2.2073	0.9833	0.4820	0.2409	0.1202
#4	9.5824	2.7178	1.2344	0.6040	0.3018	0.1506
#5	10.9749	4.2246	1.4655	0.7103	0.3573	0.1791

Table 8.2: Errors for the H-field formulation.

Chapter 9

Conclusion and outlook

9.1 Conclusion

In this work we have seen that for two-dimensional photonic crystals the Helmholtz equation can be solved numerically in a more efficient way than the divergence-type equation. This was obtained by a reformulation of the discrete equation. In the reformulated version the largest magnitude eigenvalues need to be approximated. This allows us to use standard eigensolvers with routines executing matrix-vector products, instead of routines which solve linear systems. Moreover, for the numerical approximation of the 2D Helmholtz equation with the Fourier-Galerkin method we have proved the convergence rates for the eigenvalues and eigenfunctions. This was done with similar methods as in [52]. For 3D band structure computations with the Fourier-Galerkin method we have compared the discretization of the H-field formulation with the discretization of the E-field formulation. With an example, which was also considered in [10, 39, 55, 64], we have illustrated that although the convergence rates are the same, the numerical results suggest that discretizing the E-field formulation is the better choice due to the much smaller errors for the same discretization levels. This was also observed in [64]. Recent results in [18] however, show that for finite elements the opposite is true because the piecewise regularity of the H-field is higher. Moreover, we have seen that applying the Harmonic Restarted Arnoldi method 3D band structures can be computed without any preconditioning on a desktop PC. However, there is also space for improvement in the algorithm.

9.2 Outlook

In future work we want to analyze and prove the convergence rates for the 2D divergence-type equation. Moreover, for the 3D problems the convergence with the

Fourier-Galerkin method is to be investigated. Concerning the algorithm for the numerical solution of the E-field formulation, results from the theory of structured matrices need to be observed. Inverse formulas for the inverses of multilevel BTTB matrices could improve the algorithm for the E-field formulation considerably.

Bibliography

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces. 2nd ed.* Number 140 in Pure and Applied Mathematics. Academic Press, New York, 2003.

[2] H. W. Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung. 4. Aufl.* Springer, Berlin, 2002.

[3] O. Axelsson and V. Barker. *Finite Element Solution of Boundary Value Problems (Classics in Applied Mathematics 35).* SIAM, Philadelphia, 2001.

[4] I. Babuška and J. Osborn. *Eigenvalue problems (Handbook of Numerical Analysis Vol. 2, pp. 641-787).* Elsevier North-Holland, Amsterdam and London, 1991.

[5] G. Bachman, L. Narici, and E. Beckenstein. *Fourier and Wavelet Analysis.* Springer, Berlin, 2000.

[6] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. va der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide (Software, Environments and Tools).* SIAM, Philadelphia, 1987.

[7] G. Bao, L. Cowsar, and W. Masters. *Mathematical Modeling in Optical Science (Frontiers in Applied Mathematics).* SIAM, Philadelphia (PA), 2001.

[8] D. Braess. *Finite Elemente: Theorie, Schnelle Löser und Anwendungen in der Elastizitätstheorie (Springer-Lehrbuch Masterclass).* Springer, Berlin, 2013.

[9] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations (Universitext).* Springer, New York, 2010.

[10] A. Bulovyatov. *A Parallel Multigrid Method for Band Structure Computation of 3D Photonic Crystals with Higher Order Finite Elements.* PhD thesis, Karlsruhe Institute of Technology (KIT), 2010.

[11] C. Canuto, R. H. Nochetto, and M. Verani. Adaptive Fourier-Galerkin methods. *Preprint arXiv:1201.5648 [math.NA].*

[12] R. H.-F. Chan and X.-Q. Jin. *An Introduction to Iterative Toeplitz Solvers.* SIAM, Philadelphia (PA), 2007.

[13] J. Chen, T. L. H. Li, and M. Anitescu. Parallelizing the Conjugate Gradient Algorithm for Multilevel Toeplitz Systems. *Procedia Computer Science*, 18:571–580, 2013.

[14] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems.* North-Holland, Amsterdam and New York, 1978.

[15] J. W. Cooley and J. Tukey. An algorithm for the machine calculation of complex Fourier series. *Math. of Computation*, 19.

[16] S. J. Cox and D. C. Dobson. Maximizing band gaps in two-dimensional photonic crystals. *SIAM J. Appl. Math.*, 59(6):2108–2120, 1999.

[17] S. J. Cox and D. C. Dobson. Band structure optimization of two-dimensional photonic crystals in H-polarization. *J. Comput. Phys.*, 158(2):214–224, 2000.

[18] M. Dauge, R. Norton, and R. Scheichl. Regularity of Maxwell Eigenproblems in Photonic Crystal Fibre Modelling. *Preprint arXiv:1310.7000 [math.NA].*

[19] P. J. Davis. *Circulant Matrices.* John Wiley & Sons, Hoboken (NJ), 1979.

[20] J. Demmel. *Applied Numerical Linear Algebra.* SIAM, Philadelphia (PA), 1997.

[21] D. C. Dobson. An Efficient Method for Band Structure Calculations in 2D Photonic Crystals. *J. Comput. Phys.*, 149.

[22] D. C. Dobson, J. Gopalakrishnan, and J. E. Pasciak. An Efficient Method for Band Structure Calculations in 3D Photonic Crystals. *J. Comput. Phys.*, 161(2):668–679, 2000.

[23] D. C. Dobson and J. E. Pasciak. Analysis of an algorithm for computing electromagnetic Bloch modes using Nedelec spaces. *Comput. Methods Appl. Math.*, 1(2):138–153, 2001.

[24] W. Dörfler, A. Lechleiter, M. Plum, G. Schneider, and C. Wieners. *Photonic Crystals: Mathematical Analysis and Numerical Approximation (Number 42 in Oberwolfach Seminars).* Springer, Basel, 2011.

[25] L. Evans. *Partial Differential Equations (Graduate Studies in Mathematics).* American Mathematical Society, Providence, 2010.

[26] L. Feng and V. Linetsky. Pricing options in jump-diffusion models: An extrapolation approach. *Operations Research*, 56(2):304–325, 2008.

[27] D. Fleisch. *A students guide to Maxwell's equations*. Cambridge University Press, Cambridge (UK), 2008.

[28] S. Giani and I. G. Graham. Adaptive finite flement methods for computing band gaps in photonic crytsals. *Numerische Mathematik*, 121:31–64, 2012.

[29] M. Gockenbach. *Partial Differential Equations*. SIAM, Philadelphia (PA), 2010.

[30] G. Heinig and K. Rost. Hartley transform representations of symmetric Toeplitz matrix inverses with application to fast matrix-vector multiplication. *SIAM J. Matrix Anal. Appl.*, 22(1):86–105, 2000.

[31] H. Heuser. *Lehrbuch der Analysis 1*. Vieweg Teubner, Wiesbaden, 2009.

[32] M. Hochbruck. *Skriptum zur Vorlesung Spezielle Themen der numerischen linearen Algebra SS 2013*.

[33] J. D. Jackson. *Classical Electrodynamics. 3rd Edition*. John Wiley & Sons, Hoboken (NJ), 1999.

[34] J. D. Joannopoulos, S. G. Johnson, J. N. Winn, and R. D. Meade. *Photonic Crystals: Molding the Flow of Light. 2nd Edition*. Princeton University Press, Princeton (NJ), 2008.

[35] S. Johnson and J. Joannopoulos. Block-iterative Frequency-domain Methods for Maxwell's Equations in A Planewave Basis. *Opt. Express*, 8:173–190, 2001.

[36] K. Knopp. *Theory and Application of Infinite Series*. Dover, New York, 1990.

[37] P. Kuchment. *Floquet Theory for Partial Differential Equations*. Birkhäuser, Basel, 1993.

[38] P. Kuchment. Mathematics of Photonic Crystals. In G. Bao, editor, *Mathematical Modeling in Optical Science*, pages 207–272. SIAM, Philadelphia (PA), 2001.

[39] P. Lalanne. Effective properties and band structures of lamellar subwavelength crystals: Plane-wave method revisited. *Phys. Rev. B*, 58.

[40] P. Lalanne and G. M. Morris. Highly improved convergence of the coupled-wave method for TM polarization. *J. Opt. Soc. Am. A*, 13.

[41] L. Li. Use of Fourier series in the analysis of discontinuous periodic structures. *J. Opt. Soc. Am. A*, 13.

[42] L. Li and C. W. Haggans. Convergence of the coupled-wave method for metallic lamellar diffraction gratings. *J. Opt. Soc. Am. A*, 10.

[43] X.-G. Lv and T.-Z. Huang. The inverses of block Toeplitz matrices. *Hindawi Journal of Mathematics*, 2013(Article ID 207176):8 pages, 2013.

[44] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia (PA), 2000.

[45] M. S. Min and D. Gottlieb. On the convergence of the Fourier approximation for eigenvalues and eigenfunctions of discontinuous problems. *SIAM J. Numer. Anal.*, 40(6):2254–2269, 2003.

[46] R. B. Morgan. Implicitly restarted GMRES and Arnoldi methods nonsymmetric systems of equations. *SIAM J. Matrix Anal. Appl.*, 21.

[47] R. B. Morgan. On restarting the Arnoldi method for large nonsymmetric eigenvalue problems. *Math. Comp.*, 65.

[48] R. B. Morgan and M. Zeng. Harmonic projection methods for large nonsymmetric eigenvalue problems. *Num. Lin. Alg. with Appl.*, 5.

[49] R. B. Morgan and M. Zeng. A harmonic restarted Arnoldi algorithm for calculating eigenvalues and determining multiplicity. *Linear Algebra and Its Applications*, 415(1):96–113, 2006.

[50] R. Norton. *Numerical computation of band gaps in photonic crystal fibres*. PhD thesis, University of Bath, 2008.

[51] R. Norton and R. Scheichl. Planewave Expansion Methods for Photonic Crystal Fibres. *Applied Numerical Mathematics*, 63.

[52] R. Norton and R. Scheichl. Convergence Analysis of Planewave Expansion Methods for 2D Schrödinger Operators With Discontinuous Periodic Potentials. *SIAM J. Numer. Anal.*, 47(6):4356–4380, 2010.

[53] V. Olshevsky, I. Oseledets, and E. Tyrtyshnikov. Tensor properties of multilevel Toeplitz and related matrices. *Linear Algebra and its Applications*, 412:1–21, 2006.

[54] M. Renardy and R. Rogers. *An Introduction to Partial Differential Equations (Texts in Applied Mathematics)*. Springer, New York, 2010.

[55] M. Richter. *Optimization of photonic bandgaps*. PhD thesis, Karlsruhe Institute of Technology (KIT), 2010.

[56] W. Rudin. *Principles of Mathematical Analysis*. McGraw Hill, Boston, 1976.

[57] W. Rudin. *Real and Complex Analysis*. McGraw Hill, Boston, 1987.

[58] Y. Saad. *Numerical methods for large eigenvalue problems*. Halsted Press, New York, 1992.

[59] S. Sauter and C. Schwab. *Boundary Element Methods (Springer Series in Computational Mathematics)*. Springer, Berlin, 2011.

[60] K. Schmidt and P. Kauf. Computation of the band structure of two-dimensional Photonic Crystals with hp Finite Elements. *Comp. Meth. App. Mech. Engr.*, 198:1249–1259, 2009.

[61] S. Serra Capizzano and P. Tilli. Extreme singular values and eigenvalues of non-Hermitian block Toeplitz matrices. *Journal of computational and applied mathematics*, 108:113–130, 1999.

[62] L. Shen and S. He. Analysis for the convergence problem of the plane-wave expansion method for photonic crystals. *J. Opt. Soc. Am. A*, 19(5):1021–1024, 2002.

[63] D. C. Sørensen. Implicit application of polynomial filters in a k-step Arnoldi method. *SIAM Journal on Matrix Analysis and Applications*, 13(1):357–385, 1992.

[64] H. S. Sözüer and J. W. Haus. Photonic bands: simple-cubic lattice. *J. Opt. Soc. Am. B*, 10(2):296–302, 1993.

[65] M. Stoer and R. Bulirsch. *Introduction to Numerical Analysis (Texts in Applied Mathematics)* . Springer, New York, 2002.

[66] G. Strang and G. Fix. *An Analysis of the Finite Element Method*. Wellesley Cambridge Press, Wellesley, 2008.

[67] J. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. SIAM, Philadelphia, 2004.

[68] P. Tilli. A note on the spectral distribution of Toeplitz matrices. *Linear and Multilinear Algebgra*, 45:147–159, 1998.

[69] C. van Loan. *Computational frameworks of the fast Fourier transform*. SIAM, Philadelphia (PA), 1992.

[70] A. Vretblad. *Fourier Analysis and Its Applications (Graduate Texts in Mathematics)*. Springer, New York, 2010.

[71] D. S. Watkins. *The matrix eigenvalue problem*. SIAM, Philadelphia, 2007.

[72] J. Wloka. *Partial Differential Equations*. Cambridge University Press, Cambridge (UK), 1987.

[73] A. Zygmund. *Trigonometric Series*. Cambridge University Press, Cambridge (UK), 2002.